Quantitative Data Anaylsis

Lecture notes

Instructor:
Jangho Yang
Management Sciences
Faculty of Engineering

Contents

Chapter 1: Review of Basic Probability Theory

Jangho Yang

v1.0

Contents

1 Examples of statistical thinking

1.1 Mortality rate

There is a famous study by Case & Deaton (2015) showing rising mortality of midlife white non-Hispanic men and women in the United States after 2000. Figure 1 visualizes the main result for several advanced economies where the US stands out as an exception.

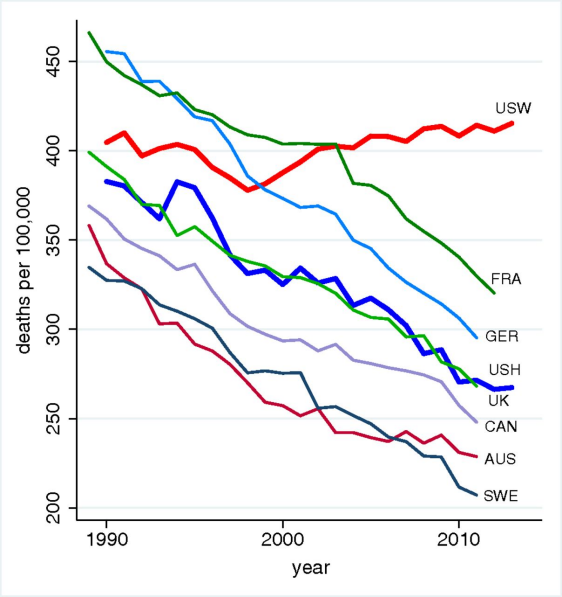


Figure 1: Figure from Case & Deaton (2015) showing rising mortality of the middle-age non-Hispanic white population in the US as opposed to the mortality trend in other countries. USW means US non-Hispanic white.

This study received significant media attention due to its inherent political implications. If no other factor is considered, the trend seems to be attributed only to the race-ethnicity. However, this important and controversial result was put under scrutiny by Gelman & Auerbach (2016) who used the same data but found that the mortality rates for the middle-aged non-Hispanic whites did not steadily increase. What they showed is that the seemingly increasing mortality rate among the middle-aged non-Hispanic whites is because the average age for this group went up, as shown in the top centered panel (B) in Figure 2. When the average age is adjusted, the mortality of midlife non-Hispanic whites doesn't show a steady increase. More importantly, the mortality rate of the non-Hispanic white males has been decreasing since 2005, while that of the non-Hispanic white females white has been dramatically increasing as shown in the bottom panel in Figure 2.

What can we learn from this example? It shows that data analysis is a subtle work whose results might change depending on which aspects of the data researchers want to reveal. Case and Deaton did not make a mistake and they correctly brought up a markedly different mortality pattern for non-Hispanic white people in the US compared to other countries. However, they used the raw data without proper age-adjustment and thus failed to separate the increasing average age from the mortality trend. In contrast, Gelman & Auerbach (2016) used the age-adjusted data and could

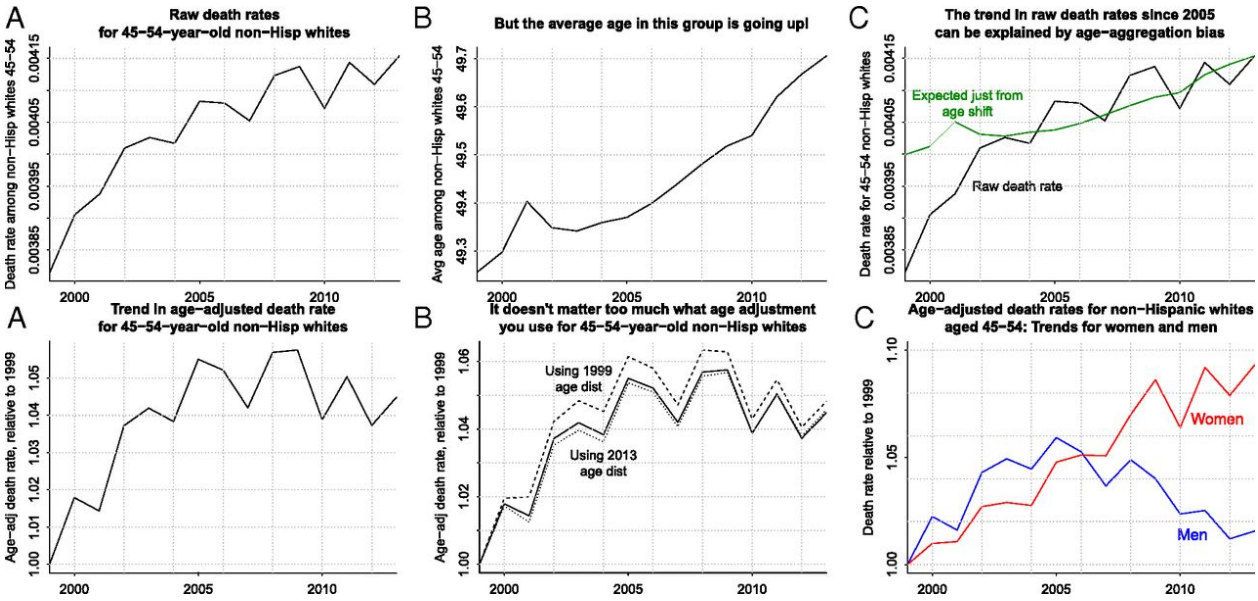


Figure 2: Figures from Gelman & Auerbach (2016) showing that the increasing mortality rate among the non-Hispanic white population is due to the rising average age in this group. When the average age is adjusted, the mortality rate for non-Hispanic white males decreases since 2015.

not only show the overall trend more correctly, but also find out the gender difference in mortality rate.

1.2 Words are new numbers

Economics is about numbers. However, there is a growing interest in doing empirical economics with words instead of numbers. For example, Thorsrud (2020) from Norges Bank constructed a public sentiment index based on the words printed in newspapers that performs well in explaining business cycles. Figure 3 showed how well the textual information contained in a daily business newspaper, matches the GDP growth in Norway.

This creative work is entirely due to a recent development in natural language processing, the methodology that has been widely used in machine learning. Without rapid progress in computational technology, this type of research would not have been possible. This is a good example of how progress in statistical techniques can give a better insight into some of the problems of central importance in our life.

1.3 Wine price and web scraping

We don't have to talk about serious stuff all the time. Sometimes, we want to drink wine with friends. Even here, statistics has something interesting to tell us. Kotonya et al. (2018) used the scraped data from the Vivino website, which provides prices, reviews & ratings of millions of world wines. For example, Figure 4 shows the price distribution by different categories. We can see that, as conventional wisdom goes, French wine and their Pinot Noir are the most expensive ones.

The key feature of this study is that the researchers did not obtain the data from a national statistical office or private consulting firm, but directly from an online website using a technique

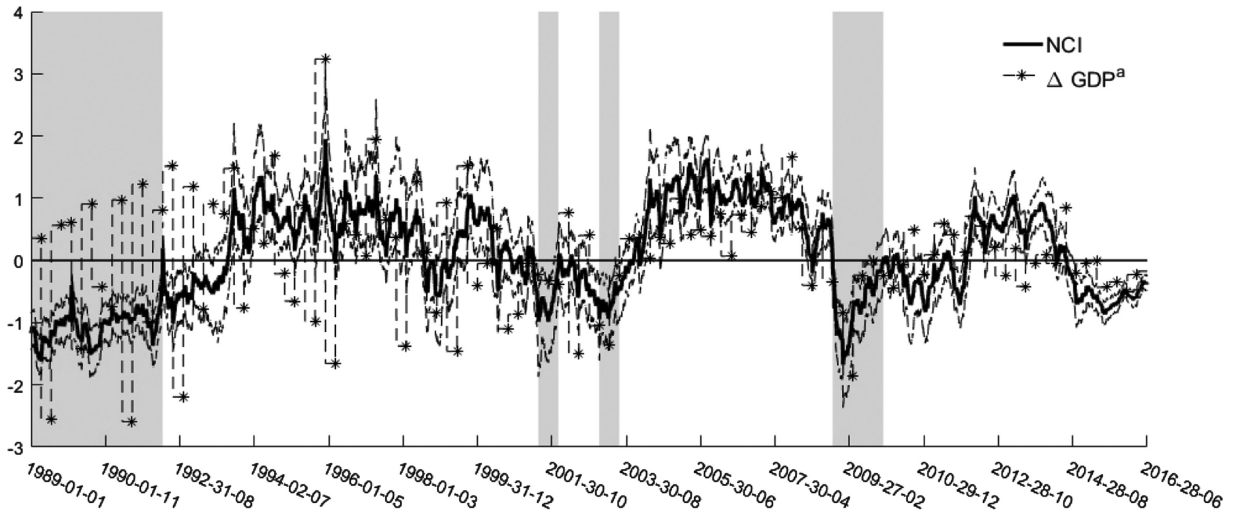


Figure 3: A figure from Thorsrud (2020) showing how textual information contained in a daily business newspaper matches well with the business cycle index based on quarterly GDP growth. NCI represents a Newsy coincident Index constructed from the words printed in newspapers.

known as web scraping or web crawling. This is also a good example of how statistical work (even descriptive one) can greatly benefit from a fast computer and efficient algorithms.

2 Statistical decision tree

The examples we discussed above utilize specific statistical tools for the purpose of a particular statistical analysis. Then, how can we choose the right statistical technique when dealing with a real-life statistical problem with data? Not surprisingly, there is a cheat sheet, known as a statistical decision tree. There are many different variations of this cheat sheet, but they follow a similar structure as shown in Figure 5.

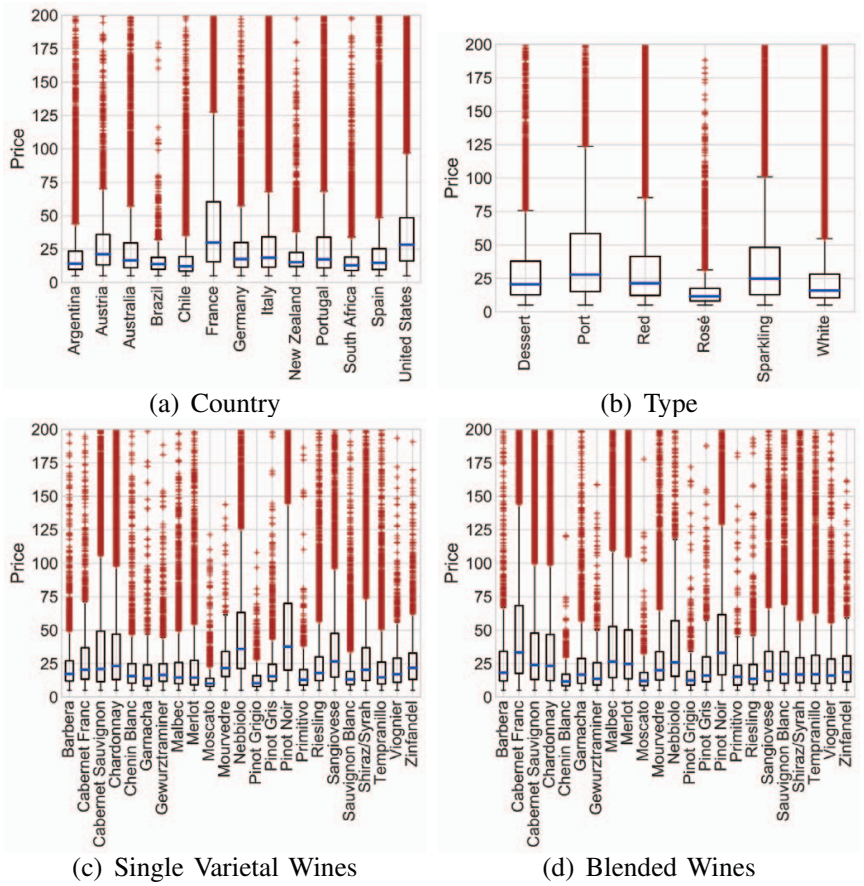
This decision tree can be helpful for finding an appropriate statistical tool among many when faced with a particular statistical problem. However, is memorizing a cheat sheet the ultimate goal of statistical learning? Obviously not. As we will discuss in this course, statistical thinking goes far beyond applying an existing toolbox mechanically. Therefore, we will forget about this decision tree for now and start from understanding a set of underlying principles of statistical thinking. By doing so, we won't have to confine ourselves to the existing toolbox when we encounter new statistical problems. Plus, we will be able to understand when the decision tree helps and when it doesn't.

To do this, we need to understand probability, the foundation of statistical inference.

3 What is probability?

3.1 A brief history

Suppose there is a bet and you can buy a ticket that guarantees you $1 when a certain outcome occurs, e.g. the head of the coin. How much are you willing to pay for the ticket? The theory of



(a) Country (b) Type

(c) Single Varietal Wines (d) Blended Wines

Figure 4: A figure from Kotonya et al. (2018) showing price distributions of wine in pounds depending on types.

probability actually originated in this practical question of how much a player should/is willing to pay for bets in gambling back in the 17th century. The probability concept was understood as a tool for quantifying a *subjective* chance of a betting outcome. It was subjective since the probability was associated with the opinion or willingness of a better. Since the theory of probability developed as a theory of chances or a theory of betting, statisticians/philosophers such as Thomas Bayes and Pierre-Simon Laplace back then contemplated on seemingly unanswerable questions using the theory of probability such as the chance of sun rising the next day.[1]

Starting in the 19th century, the concept of probability took a radical turn when a group of statisticians (Karl Pearson, and R. A. Fisher) began to use probability as observed *frequencies*. It was radical because they understood probability as being *objective* in the sense that it refers to the frequency of an event occurring over an infinite number of observations generated from a *true process*. That is, these statisticians assumed that there is a true data-generating process, which is unknown to us, but reveals itself in constant frequencies of various outcomes in infinitely long series of observations.

[1]This is a very interesting statistical problem. We will come back to this question when we discuss Laplace's rule of succession.

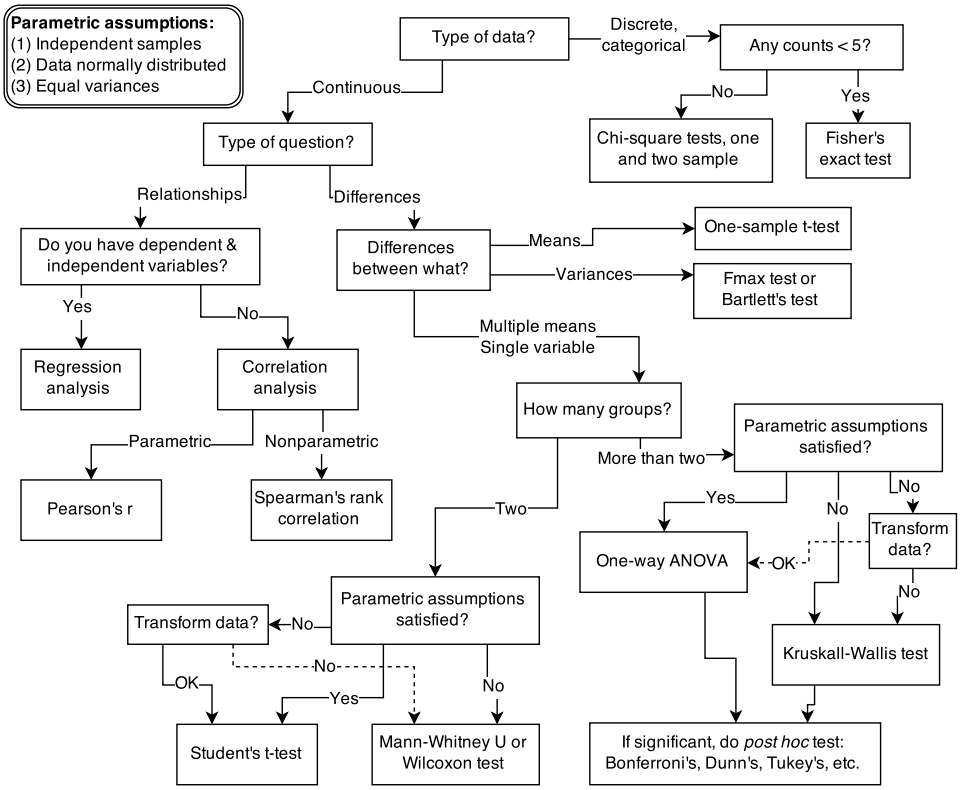


Figure 5: A statistical decision tree from McElreath (2020).

The history of statistics in the 20th century was punctuated by dialogues/clashes between these two different schools of thoughts, namely, Bayesians vs. Frequentists. This course does not subscribe to one of these two different theories. Instead, the main objective of the course is to give students the foundation for understanding and actively engaging with statistical problems from a wide range of methodological perspectives and, therefore, we will introduce and compares various statistical approaches, including both Bayesian and Classical.

3.2 Concept

With this philosophical debate in mind, let's formally define the probability that is applicable (at least loosely) to the different methodologies that we will be covering throughout this course.

Suppose we have a repeatable and countable *experiment* such as a coin toss, a clinical trial of a drug, or drawing a ball from an urn of balls with red, blue, and green colors. Each experiment has a set of possible *outcomes*. For example, a head or a tail in a coin tossing, ameliorating vs. deteriorating effects in a drug test, and a red, blue or green ball in the urn example. Now, we define *sample space* (or *event space*) Ω as the set of all possible outcomes of an experiment. We also define *event* as a subset of a sample space, $e_1, e_2, \ldots, e_k$.

A naive/primitive definition of *probability* of the event e_i, $p(e_i)$ is the number of favorable outcomes in e_i over the number of all possible outcomes in Ω.

$$P(e_i) = \frac{\text{\# of favorable outcomes in } e_i}{\text{\# of all possible outcomes in } \Omega} \tag{1}$$

Note that this is a naive/primitive definition because we need to assume that all outcomes are equally likely and that the sample space is finite. A probability space or a system of probabilities over the sample space Ω is a list of non-negative $P(e_i)$ that, by definition, add up to 1.

$$\sum_{i=1}^{k} P_i = 1, \quad i = 1, \ldots, k \tag{2}$$

It is helpful to think of probability as a function that takes an event e_i (a subset of Ω) as input and gives some number between 0 and 1 as output. Note that since calculating probability involves calculating the number of favorable outcomes, combinatorics becomes really important.

Examples

Let's take some simple examples. Suppose we conduct the experiment of tossing a coin twice. The outcome of each experiment is having a tail (T) or an head (H). What is the probability of seeing a head at least once? In this experiment, there are a total of 4 possible outcomes, HH, HT, TH and TT, where H means a head and T means a tail. The number of favorable outcomes in the event of our interest (# of heads ≥ 1) are 3 because all HH, HT, and TH have H at least once. Therefore, the probability of seeing a head at least once in our coin toss experiment is 3/4.

Second, let's throw two dice. What is the probability that the total is 10? The possible favorable outcomes are {6,4}, {4,6}, and {5,5}. Since the number of all possible outcomes is 36 ($36 = 6^2$, 6 faces of the die thrown 2 times), the probability that the total is 10 is 3/36.

3.3 Basic properties and notation

There are some important properties of probability that are conceptually important and will help us to calculate some complicated probabilities later on as well. I will use Venn diagrams to illustrate some of the properties (See Figure 6).

1. Suppose event A and B. By definition, we have

$$P(A) = \sum_{s_i \in A} P(s_i)$$
$$P(B) = \sum_{s_i \in B} P(s_i)$$

where s_i is a favorable outcome in a relevant event.

2. The probability of an empty set (zero outcomes) is zero $P(\emptyset) = 0$ while the probability of the entire set (all possible outcomes) $P(\Omega) = 1$.

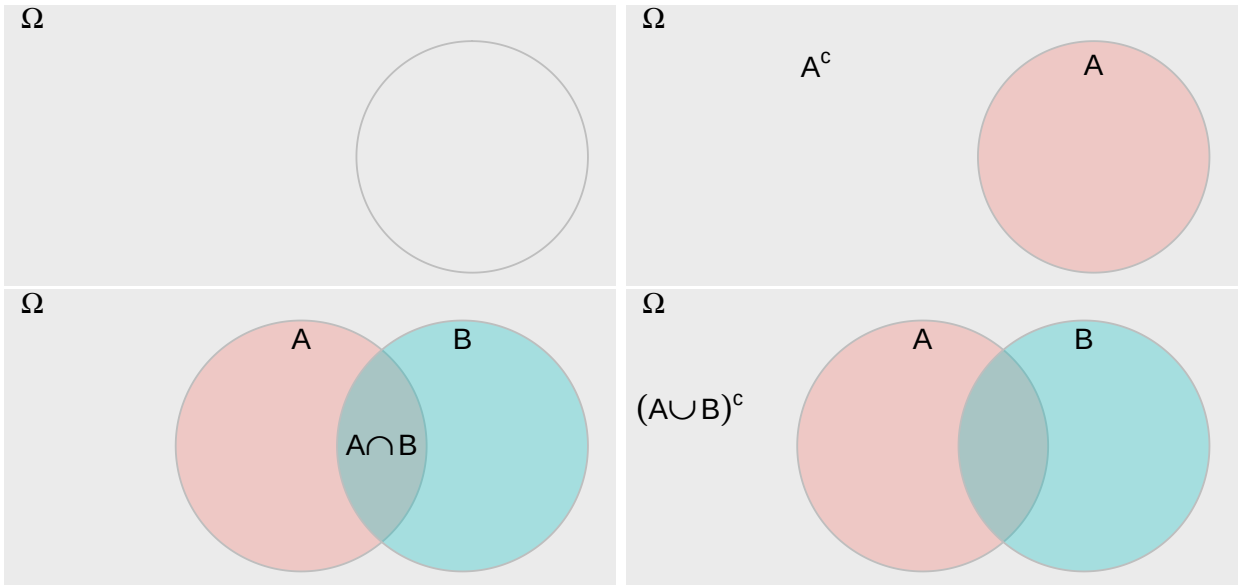


Figure 6: The entire sample space Ω and its two subsets A and B.

3. The complement of an event A is A^c: $P(A^c) = 1 - P(A)$

4. The probability that both A and B occur, i.e., the joint probability of A and B (intersection of events A and B): $P(A \cap B)$.

5. The probability that at least one of the event occurs (union of events A and B): $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

6. The probability that none of A and B occurs: $P(A \cup B)^c = 1 - P(A \cup B)$

3.4 Conditional probability and independence

Conditional probability
As we can see in the case of the union and intersection of events, the probability can be defined for multiple events as well. In this case, it is important to understand how each event is related to one another. One key question to ask is whether the occurrence of the event affects the occurrence of other events. For example, drawing a red ball without replacement from an urn with red and white balls affects the chance of drawing another red ball subsequently since the first draw changes the number and the composition of balls in the urn. In contrast, when drawing a red from the same urn with replacement, meaning that we put the ball back when we draw it, the chances of drawing another red ball do not change. This example shows that we need to understand how each event is related to others when dealing with multiple events.

One key concept that captures the relationships between the occurrences of events is the conditional probability. For events A and B, the conditional probability is defined as the probability that A occurs given that B occurs

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \tag{3}$$

Going back to the Venn diagram above, the conditional probability is the fraction of the event B where both A and B occur. Using a more technical word, the probability space is *normalized* with respect to event B, which explains why we have the probability of B in the denominator. See Figure 7 for visualization.

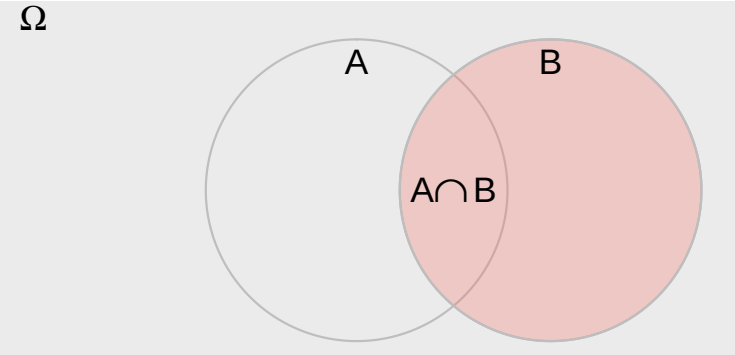


Figure 7: Normalization with respect to set A.

Independence

Like drawing a ball from an urn without replacement, there are many cases when the occurrence of one event affects the occurrence of other events. For example, the probability of getting admitted to a university *depends* on which department/school one applies to. In this case, we say that events are *dependent*. When we have dependent events, the conditional probability $P(A|B)$ differs from the *unconditional* or *marginal probability* $P(A)$ since the occurrence of event B affects the occurrence of event A.

What happens if event B doesn't have any information about event A? That is, if the occurrence of event B does not affect the probability of occurrence of event A, we say events A and B are *independent*. In this case, the conditional probability $P(A|B)$ is the same as the unconditional or *marginal probability* $P(A)$.

3.5 Joint probability

From the Equation 3, the joint probability of A and B can be calculated with a simple multiplication rule as follows

$$\begin{aligned} P(A \cap B) &= P(A|B)P(B) \quad \text{or} && (4) \\ P(A \cap B) &= P(B \cap A) && (5) \\ &= P(B|A)P(A) && (6) \end{aligned}$$

Note that when events A and B are independent, the joint probability simply becomes a product of two marginal probabilities[2], .

3.6 Examples

Urn example

Now let's look at a simple urn example. Suppose we have one urn with 5 black balls and 3 white balls. We draw 3 balls *without replacement*, meaning that when we draw a ball we do not put it back. Then, what is the probability of drawing 2 black balls and 1 white ball? The most tedious way of calculating this probability is to find each probability of $\{B, B, W\}$, $\{B, W, B\}$ and $\{W, B, B\}$, which are the cases of two balls out of three, and add them together. We can easily see from the definition of the joint distribution in Equation 6 that the probability of $\{B, B, W\}$ is $P(X_1 = B, X_2 = B, X_3 = W) = P(X_1 = B)P(X_2 = B|X_1 = B)P(X_3 = W|X_1 = B, X_2 = B)$ = (5 black balls / 8 remaining balls)*(4 black balls / 7 remaining balls)*(3 white balls / 6 remaining balls), where $X_.$ represents the color of the balls. Note that the probability of drawing the second black ball is conditional on the draw of the first black ball since we are drawing a ball without replacement. Since the probability of two black balls is the sum of these three cases $\{B, B, W\}$, $\{B, W, B\}$ and $\{W, B, B\}$, the answer is $(5/8) * (4/7) * (3/6) + (5/8) * (3/7) * (4/6) + (3/8) * (5/7) * (4/6) = 0.536$.

There is another way of calculating this probability using a simple combination rule, a counting technique when the order does not matter. Refer to Appendix A.1 for an overview of combinations (or Binomial coefficient).[3] Using combinations, the possible favorable outcomes for two black balls and one white ball can be calculated simply by $\binom{5}{2} \times \binom{3}{1}$, meaning that we choose 2 black balls out of 5 and choose 1 ball from 3 white balls. The number of all possible outcomes of drawing 3 balls is $\binom{8}{3}$, meaning that we choose any 3 balls from all 8 balls. Therefore, the probability of drawing 2 black balls and 1 white ball is $\binom{5}{2} \times \binom{3}{1} / \binom{8}{3} = 0.536$.

Birthday problem

Let's looks at another problem, namely a birthday problem. Suppose we have a group of 10 people. Then, what is the probability that at least two people among this group have the same birthday? Here, we will use the complement rule and calculate the probability of no match first and then subtract it from 1, $1 - P(\text{no match})$. The number of favorable outcomes in no match is simply $365 \times 364 \times \cdots \times 356$. That is, the first person's birthday is anything from 365 days, and the second person's birthday is anything but the first person's birthday, which is any day from 364

[2]Note that the notation $P(A \cap B)$ and $P(A, B)$ are equivalent.

[3]Combinations, nC_k, and the binomial coefficient, $\binom{n}{k}$, are mathematically the same: $nC_k = \binom{n}{k} = \frac{n!}{k!(n-k)!}$. We will use the notation for the binomial coefficient throughout this course.

days, and the third person's birthday is anything but the first and the second person's birthdays, which is any day from 363 days, and so forth. Since the number of all possible outcomes for 10 persons' birthday is 365^{10}, we can calculate the probability that at least two people among this group have the same birthday as

$$P(\text{match}) = 1 - \frac{365 \times 1 \times \cdots \times 356}{365^{10}} = 0.117 \tag{7}$$

That is, there is an 11.7% chance that at least two people among a group of 10 people have the same birthday.

Newton–Pepys problem

Finally, let's throw multiple fair dice and calculate the probability of observing "6".

1) Suppose we throw 6 dice. What is the probability that at least one "6" appears?

$$P(A) = 1 - \left(\frac{5}{6}\right)^6 \approx 0.66$$

2) 12 dice: at least two "6" appears.

$$P(B) = 1 - \sum_{x=0}^{1} \binom{12}{x} \left(\frac{1}{6}\right)^x \left(\frac{5}{6}\right)^{12-x} \approx 0.62$$

3) 18 dice: at least three "6" appears.

$$P(C) = 1 - \sum_{x=0}^{2} \binom{18}{x} \left(\frac{1}{6}\right)^x \left(\frac{5}{6}\right)^{18-x} \approx 0.60$$

As we can see, these seemingly same exercises have different probabilities. Interestingly enough, the first case has the highest probability. This is exercise is called Newton–Pepys problem. You can refer to Stigler (2006) for more technical details about this problem.

3.7 Bayes' Theorem

From the simple multiplication rule in Equation 6, we can derive a very important theorem in probability theory, called Bayes' Theorem. Since $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$, we have

$$P(A|B) \quad = \quad \frac{P(B|A)P(A)}{P(B)}. \tag{8}$$

Note that $P(B) = \sum_{k \in A} P(B|A = k)P(A = k)$ (the law of the total probability). This simple relation has a far-reaching implication in statistical inference, which will be discussed in detail in Topic 3. For now, let's try to understand what this theorem means using some examples.

Monty Hall problem

There was a television game show called "Let's Make a Deal." The rule of the game goes like this. There are three boxes where one box contains a great prize and two boxes are empty. The contestant chooses one box first, and then, the host (Monty) opens one of the two other empty

boxes. After this, the host asks the contestant if he/she wants to switch from the chosen box to the remaining one. Should the contestant switch or stick to the original box? This seemingly simple exercise embodies the gist of Bayes' theorem that a probability decision should be conditioned on whatever information available. Let's do some calculations and see what the answer is.

To solve this problem, we need to define two random variables: X_i for $i = 1, 2, 3$ is the event that the prize is in the box i. K_i for $i = 1, 2, 3$ is the event that the host opens the box i. Without loss of generality, we will suppose that the contestant chose Box 1 and the host opened Box 2. Then, the question boils down to calculating the probability of the prize being in Box 3 conditional on that the host opened Box 2, $P(X_3|K_2)$. Using Bayes' Theorem, we can show

$$P(X_3|K_2) = \frac{P(K_2|X_3)P(X_3)}{P(K_2|X_1)P(X_1) + P(K_2|X_2)P(X_2) + P(K_2|X_3)P(X_3)} \tag{9}$$

Let's unpack the right-hand-side one by one. First, before this game starts, each box has an equal probability of containing the prize. If the condition is violated, the game is not fair. Therefore, we have:

$$P(X_1) = P(X_2) = P(X_3) = 1/3$$

$P(K_2|X_3)$ is the probability of the host opening Box 2 when the prize is in Box 3 (and the contestant chose Box 1). Since the rules of the game mandates the host to open the empty box, Box 2 will be opened when the host knows that Box 3 has the prize. Therefore

$$P(K_2|X_3) = 1$$

$P(K_2|X_1)$ is the probability of the host opening Box 2 when the prize is in Box 1. The host can choose either Box 2 and Box 3 when the contestant chose Box 1 with the prize in it. Because there is no reason to believe a priori that the host is biased to either of these two boxes, the probability of the host opening Box 2 is 1/2.

$$P(K_2|X_2) = 1/2$$

Finally, $P(K_2|X_2)$ is the probability of the host opening Box 2 when the prize is in Box 2. Again, the host is mandated to open the empty box. So this probability is zero.

$$P(K_2|X_2) = 0$$

With this, we can calculate $P(X_3|K_2)$ as follows:

$$P(X_3|K_2) = \frac{1 \times \frac{1}{3}}{\frac{1}{2} \times \frac{1}{3} + 0 \times \frac{1}{3} + 1 \times \frac{1}{3}} = \frac{2}{3}$$

Since the probability of the prize being inside Box 3 is 2/3 and is greater than 1/2, the contestant should switch to Box 3. To get a bit of intuition behind this puzzle, suppose we have 1 million boxes instead of 3 boxes. As before, the contestant opens the first box and the host opens all the empty boxes until there is only one box left. Do you want to switch to the last remaining box? Probably yes because the chance you initially picked the box with the prize is only 1/1,000,000, while the chance the remaining box is the right one should be higher than 1/1,000,000 after the host removing all of the empty boxes for the contestant.

Medical diagnosis

Suppose there is a patient who has been tested positive for a very rare disease that only appears in 0.01% of the population. However, the test is not perfect and is known to have a 2% false-positive rate and a 1% false-negative rate. With this information, what is the probability that the patient has the disease?

As above, we need to define two random variables: D is the event that the person has the rare disease and T is the event that the test for the disease is positive. What we are interested in is the probability of the person having the disease conditional on the test coming out positive, $P(D|T)$. Using Bayes' Theorem, we have

$$P(D|T) = \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|\bar{D})P(\bar{D})} \tag{10}$$

where $P(\bar{D})$ is $1 - P(D)$, the event that the person does not have the disease. $P(D)$ is the prior (or marginal) probability of the disease, which is 0.01, so

$$P(D) = 0.0001 \quad P(\bar{D}) = 0.9999 \tag{11}$$

$P(T|D)$ is the probability of a positive test conditional on the person having the disease. Since the false-negative rate is 1%, meaning that the test identifies the patient with the true disease 99%, we have

$$P(T|D) = 0.99 \tag{12}$$

Finally, $P(T|\bar{D})$ is the probability of a positive test conditional on the person not having the disease. Since the false positive rate is 2%, we have

$$P(T|\bar{D}) = 0.02 \tag{13}$$

Therefore, we can calculate $P(D|T)$ as follows

$$P(D|T) = \frac{0.99 \times 0.0001}{0.99 \times 0.0001 + 0.02 \times 0.9999} = 0.00492 \tag{14}$$

Therefore, the probability of the person with the positive test actually having the rare disease is 0.492%. The reason for such a low probability is that the prior probability of the disease is extremely low.

4 Probability distributions

4.1 Random variable

So far, we have focused on the probability of a single event. Now, we can think of assigning a probability to each of all possible events in Ω. To do this, we first need a function that relates sample space Ω to the real line, which we call the *random variable*. This random variable is necessary because while the sample space can consist of anything (even non-numbers such as hit or miss), the probability space always needs to be within the real line between 0 and 1. For example, if the sample data consists of observations on a sequence of coin tossing, a random variable X takes two values: 1 (a "head") or 0 (a "tail"). In this sense, a random variable is the

realization of the random phenomena from the sample space Ω as real numbers.

Let's take another example. Suppose that we are interested in how often we hear the word "probability" in a history class. The sample space in this case already consists of integers from 0 to infinity and therefore the random variable takes values from 0 to infinity.

Discrete and continuous random variable

Here we divide the random variables into two groups: *discrete* and *continuous* random variable. Loosely speaking, discrete values are enumerable values that can be determined accurately while continuous values are non-enumerable that cannot be determined accurately due to the fact that there are infinitely many values given the interval. Examples of discrete variables include heads or tails in a coin tossing and the number of babies born each year since all can be translated to integer numbers. Examples of continuous variables include waiting time and heights, which can be translated into real numbers.

4.2 Probability functions

Probability density/mass function

Now that we established the concept of the random variable, let's define a function that assigns probabilities to each possible value of a random variable X. Here, we introduce the *probability mass function* (PMF) and the *probability density function* (PDF). The PMF is a function that assigns probabilities to each possible value of a discrete random variable X, which we will denote by p_X. By definition,

$$p_X(x) \equiv P(X = x) \tag{15}$$

where $\sum_{x\in\Omega} = 1$. The PDF is the probability function but is defined for a continuous random variable X. Note that it is "density," which is only defined given unit volume or area:

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx \tag{16}$$

where $\int_{-\infty}^{\infty} f_X(x)dx = 1$. This implies that the PDF assigns a probability $f_X(x)dx$ to the interval $[x, x + dx]$. We will discuss some examples of PDFs and PMFs in Section 6.

Cumulative probability distribution

The *cumulative probability distribution* (CDF) is a function that gives the probability that the random variable X takes a value less than or equal to x:

$$F_X(x) = P(X \leq x) \tag{17}$$

A valid CDF is a monotonically non-decreasing function within an interval of 0 and 1. It is worthwhile to mention that all PMFs and PDFs can be derived from CDFs. This is because probability functions can be understood as changes in the CDF.

4.3 Joint, conditional, marginal distribution

Based on the basic building blocks of probability functions reviewed in Section 3, we can define three important types of probability distributions.

First, the *joint probability distribution*. When random variables have more than one dimension, we can construct a joint probability distribution. Suppose we have two discrete random variables X and Y. The joint probability distribution of X and Y is:

$$p(x, y) = P(X = x, Y = y) \tag{18}$$

Conditional probability distribution

Second, the *conditional probability distribution*. Suppose we have two discrete random variables, X and Y. The conditional distribution of X is the probability distribution of X holding constant Y at some particular value.

$$p(x|y) = P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} \tag{19}$$

Marginal probability distribution

Third, the *marginal probability distribution*. Given the joint distribution, the marginal probability distribution is the probability distribution of a random variable irrespective of other random variables. Marginalization means "aggregation" in probability theory which effectively makes the distribution unconditional. Suppose we have two discrete random variables, X and Y, and the joint distribution $P(X, Y)$, the marginal distribution of X is:

$$p(x) = \sum_y P(X = x, Y = y) = \sum_y P(X = x|Y = y)P(Y = y) \tag{20}$$

Note that the marginalization for joint PDF requires integral, $f(x) = \int_{-\infty}^{\infty} f(x, y)dy$. When the random variables X and Y are independent, we can show that the joint distribution of X and Y is the product of the marginal distributions of X and Y.

$$p(x, y) = p(x)p(y) \tag{21}$$

5 Characterizing probability distributions

So far, we have discussed different types of probability distributions. We now turn our attention to how to summarize the probability distribution.

5.1 Expected value

The *expected value* of a discrete random variable X is defined as

$$\mathrm{E}(X) = \sum_x xp(x) \tag{22}$$

The expected value is essentially the weighted average of all realizations of a random variable, weighted by their probabilities (relative occurrence). Note that we need to replace summation with integral in the case of the expected value of a continuous random variable: $\mathrm{E}(X) = \int_{-\infty}^{\infty} xf(x)dx$. Importantly, the expected value is a linear operator. That is, the expected value of the sum of random variables is equal to the sum of their individual expected values and the expected value scales linearly with a multiplicative constant:

$$\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y) \tag{23}$$

$$\mathrm{E}(cX) = c\,\mathrm{E}(X) \tag{24}$$

A classic example of the expected value is the St. Petersburg Paradox. Suppose we toss a coin infinitely. If a head is landed k times in a row, you're given 2^k dollars. Let X represent the expected value of a dollar earned for this game. Then,

$$\begin{aligned} E(X) &= \frac{1}{2} \cdot 2 + \frac{1}{4} \cdot 4 + \frac{1}{8} \cdot 8 + \frac{1}{16} \cdot 16 + \cdots \\ &= \sum_{k=1}^{\infty} 2^k \times \frac{1}{2^k} \\ &= +\infty \end{aligned}$$

The expected value of this gambling is infinite. How much money would you bet?

5.2 Variance

The *variance* of a random variable X is defined as

$$\mathrm{Var}(X) = \mathrm{E}((X - E(X))^2) \tag{25}$$

The variance gives a degree of dispersion/variability of the random variable X by measuring how much observations are spread out from their expected value. A standard deviation of X, σ_X is defied as $\sqrt{\mathrm{Var}(X)}$, which gives a measure of dispersion in the units of the random variable.

When we have multiple random variables, we can measure the joint dispersion/variability of any pair of them. Suppose we have two random variables X and Y. Then, the *covariance* of X and Y is:

$$\mathrm{Cov}(X, Y) = \mathrm{E}(X - E(X))\,\mathrm{E}(Y - E(Y)) \tag{26}$$

How to interpret the sign and the magnitude of the covariance of X and Y? If positive, X's variability is positively related to Y's variability, meaning that X and Y move in the same direction, e.g. X increases (decreases), then Y increases (decreases). If negative, X and Y move in the opposite direction. The magnitude shows the degree of such co-movement of X and Y. If zero, this means that X and Y do not move together.

5.3 Covariance and correlation

Covariance is not comparable across different sets of random variables. e.g. comparing the covariance of height and weight with the covariance of exercise and calorie burning. They can't be comparable due to the different units used in each variable. To do this, we need the covariance of normalized/unit-free random variables. This normalized covariance is called *correlation* and is defined as:

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} \tag{27}$$

$$= \mathrm{Cov}\left(\frac{X - E(X)}{\sigma_X}, \frac{Y - E(Y)}{\sigma_Y}\right) \tag{28}$$

Note that $(X - E(X))/\sigma_X$ is a handy standardization operator. It is standardization because the mean of this operator is always 0 and the variance (and the standard deviation) is always 1.

Since we measure joint variability of two standardized variables whose variance is 1, correlation is always bounded between -1 and 1. (See Appendix A.2 for proof.)

Unlike covariance, correlation is comparable across different sets of random variables. Suppose that the correlation between height and weight is 0.4, while the correlation between exercise and calorie burning is 0.9. Since the correlation index is unit free, we can say that exercise and calorie burning move much more in tandem compared to height and weight variables.

6 Well-known probability distributions

There are a number of well-established probability distributions, most of which will not be discussed in this section. I will provide a detailed discussion of some distributions when we need to use them later in the course, e.g. the Poisson distribution for Poisson regression. In this section, we will discuss only several distributions briefly to give students a sense of how probability distributions can be constructed from the underlying data-generating process and how they behave differently with varying parameter values.

6.1 Discrete probability distribution

Binomial distribution

The binomial distribution is the probability distribution of n repeated Bernoulli trials with success probability p, e.g. how many heads when tossing a coin n times. In this setting, a random variable X represents k number of success, $k = 0, \ldots, n$. We can show that the probability distribution takes the form of

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{29}$$

The expected value and the variance are $\mathrm{E}[X] = np$ and $\mathrm{Var}(X) = np(1-p)$, respectively. See Figure 8.

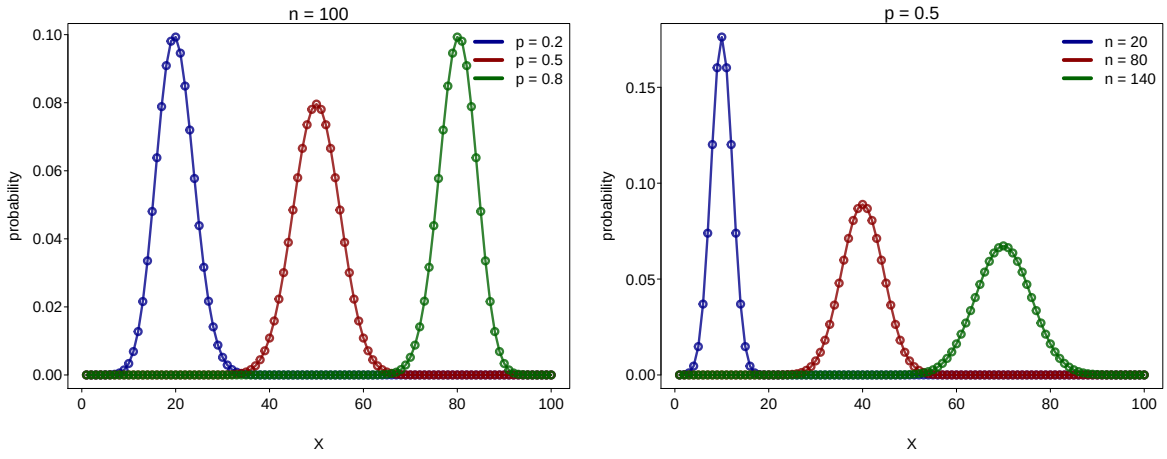


Figure 8: Binomial probability mass function with varying parameters

Geometric distribution

Another simple extension of the Binomial distribution is the *Geometric distribution* whose underlying process is the same repeated Bernoulli trials with the success probability p. In the Geometric distribution setting, what we are interested in is not the number of success out of n trials, but the number of failures before the first success, e.g. the number of tails before the first head when tossing a coin multiple times. [4] A random variable X represents k number of failures for $k = 0, 1, 2, \ldots$. The probability distribution takes the following form:

$$P(X = k) = (1 - p)^k p \tag{30}$$

The expected value and the variance are $\mathrm{E}[X] = 1 - p/p$ and $\mathrm{Var}(X) = 1 - p/p^2$, respectively. See Figure 9.

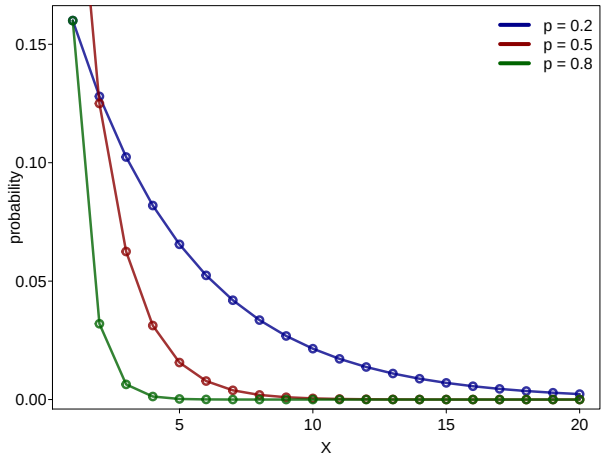


Figure 9: Geometric probability mass function with varying parameters

Negative binomial distribution

An extension of the Geometric distribution is the *Negative Binomial distribution* whose underlying process is the same repeated failures before the success with the success probability p. In the negative Binomial setting, we repeat the same process underlying the geometric distribution and look at the number of failed trials before r-th success, e.g. how many tails before two heads (r=2) when tossing a coin multiple times. A random variable X represents the k number of failures, $k = 0, 1, 2, \ldots$ given r-th success. The probability distribution takes the following form.

$$\Pr(X = k) = \binom{k + r - 1}{r - 1} (1 - p)^k p^r \tag{31}$$

The expected value and the variance are $\mathrm{E}[X] = rp/(1-p)$ and $\mathrm{Var}(X) = pr/(1-p)^2$, respectively. See Figure 10.

6.2 Continuous probability distribution

Exponential distribution

[4] By symmetry, the number of success before the first failure gives the same distribution. For notational consistency, we choose to use the number of failures before the first success.

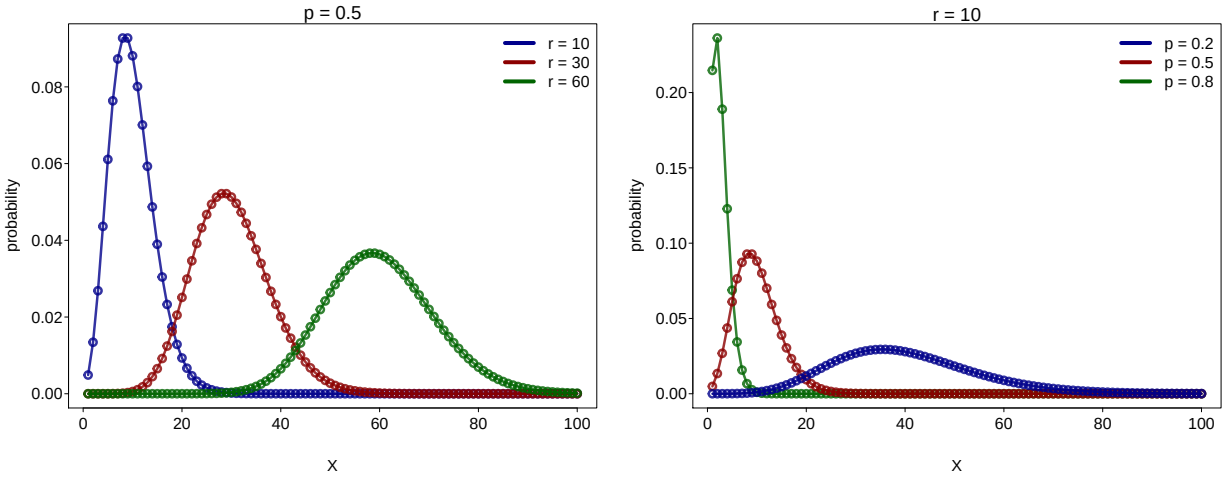


Figure 10: Negative binomial probability mass function with varying parameters

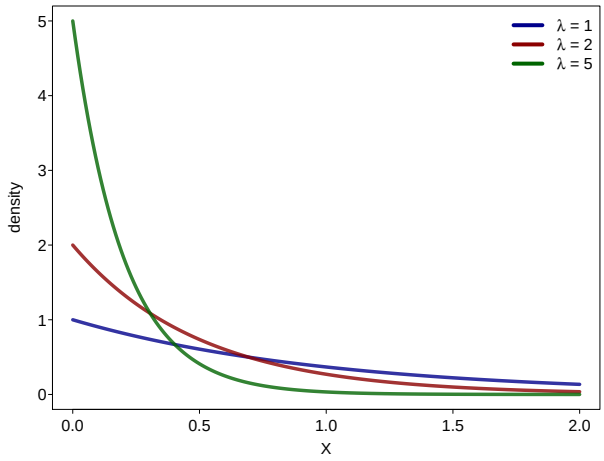


Figure 11: Exponential probability density function with varying parameters

The *exponential distribution* is the continuous analog of the geometric distribution. Here, instead of the number of failure before the first success, we are interested in the duration of waiting time before the first event occurring given the"rate" parameter λ representing the unit frequency of events occurring.[5] A random variable X represents the inter-arrival time x given λ:

$$f(x) = \lambda e^{-\lambda x} \tag{32}$$

The expected value and the variance are $\mathrm{E}[X] = 1/\lambda$ and $\mathrm{Var}(X) = 1/\lambda^2$, respectively. See Figure 11.

Gamma distribution

The *Gamma distribution* is the continuous analogue of the negative-binomial distribution. Here, instead of the number of failure before the r-th success, we are interested in the waiting time before

[5] As we will see later in the course, the parameter λ comes from the mean rate of the Poisson process.

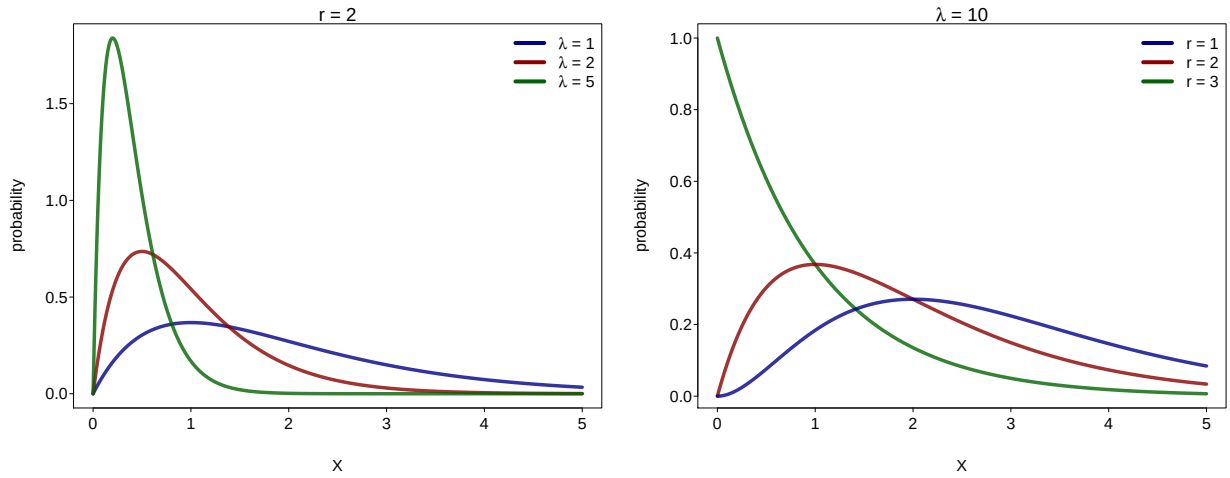


Figure 12: Gamma probability density function with varying parameters

the r-th success given the same λ parameter as in the exponential case. A random variable X represents the inter-arrival time x before r-th event given λ:

$$f(x; r, \lambda) \quad = \frac{\lambda^r x^{r-1} e^{-\lambda x}}{\Gamma(r)} \qquad \text{for } x > 0 \quad r, \lambda > 0, \tag{33}$$

where r is often called *shape parameter* and $\Gamma(.)$ is a gamma function: $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx$.. When $r = 1$, the Gamma distribution is equivalent to the exponential distribution. The expected value and the variance are $\mathrm{E}[X] = r/\lambda$ and $\mathrm{Var}(X) = r/\lambda^2$, respectively. See Figure 12.

Normal distribution

The *Normal distribution* is a distribution of the aggregate sum (or the average) of the random variables. That is, when we draw n random numbers from any distributions and take the average of them (and we repeat this many times), this average will converge to a Normal distribution. This important property can be proven by the *Central Limit Theorem*, which will be sketched out later. The Normal distribution has the following PDF

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{34}$$

The expected value and the variance are $\mathrm{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$, respectively. See Figure 13.

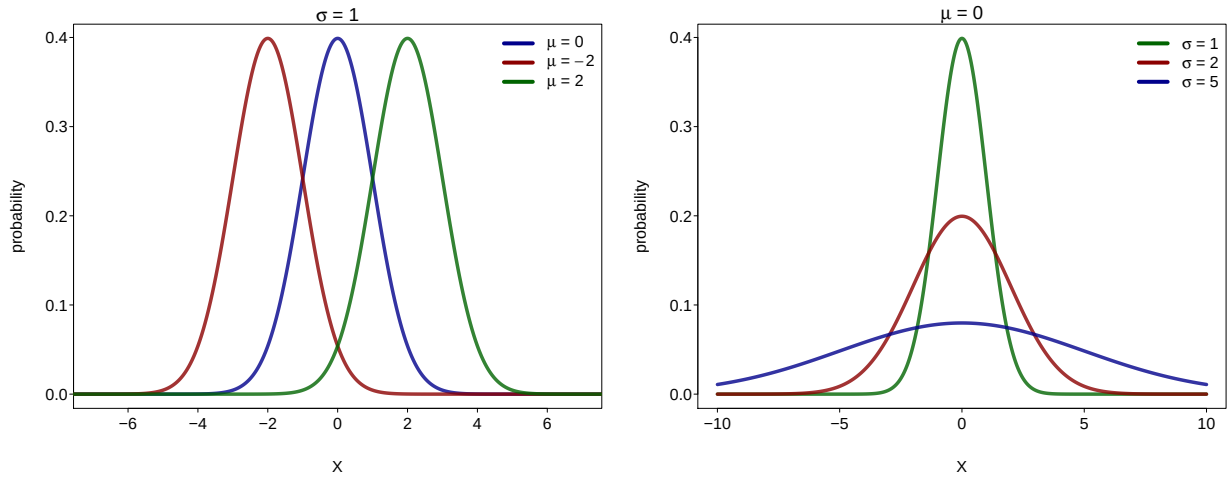


Figure 13: Normal (Gaussian) probability density function with varying parameters

A Some useful math

A.1 Binomial coefficient

The binomial coefficient refers to $\frac{n!}{(n-k)!k!} = \binom{n}{k}$, which is often read as "n choose k" and is called the choose function of n and k.

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!} = \frac{n \cdot (n-1) \cdots (n-k+1)}{k \cdot (k-1) \cdots 1} \quad \text{if } k \in \{0, 1, \ldots, n\}, \qquad (1)$$

and,

$$\binom{n}{k} = 0 \quad \text{if } k > n.$$

This represents the number of ways that k elements can be chosen from among n elements, when order is irrelevant, that is, the possible combinations of k length elements from a set with n elements without considering the ordering of the combination. It can be also interpreted as the number of different paths in the tree diagram for an n-sequence of Bernoulli trials for which the number of successes is k.

A.2 Bounds for the Correlation Coefficient

Cauchy–Schwarz inequality implies that

$$|\operatorname{E}(XY)|^2 \leq \operatorname{E}(X^2)\operatorname{E}(Y^2). \qquad (35)$$

Using this inequality, we can prove $\rho(X, Y) \leq 1$. Let μ_x and μ_y be the mean of X and Y, respectively. Then,

$$\begin{aligned} \operatorname{Cov}(X,Y)^2 &= [\,E((X-\mu_x)(Y-\mu_y))]^2 \\ &\leq \operatorname{E}((X-\mu_x)^2)\operatorname{E}((Y-\mu_y)^2) \\ &= \operatorname{Var}(X)\operatorname{Var}(Y) \end{aligned}$$

Therefore, $\frac{\operatorname{Cov}(X,Y)^2}{\operatorname{Var}(X)\operatorname{Var}(Y)} \leq 1$, and

$$-1 \leq \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}} = \rho(X,Y) \leq 1 \qquad (36)$$

A.3 Adam's law and Eve's law

Adam's law: Law of total expectation

The expected value of unconditional a random variable X is the conditional expectation of X on another random variable Y:

$$\operatorname{E}(X) = \operatorname{E}(\operatorname{E}(X \mid Y)) \qquad (37)$$

Eve's law: Law of total variance

The variance of a random variable X is the sum of the expected value of the conditional variance of X on Y and the variance of the conditional expectation of X on Y:

$$\operatorname{Var}(X) = \operatorname{E}[\operatorname{Var}(X \mid Y)] + \operatorname{Var}(\operatorname{E}[X \mid Y]). \qquad (38)$$

References

Case, A. & Deaton, A. (2015), 'Rising morbidity and mortality in midlife among white non-hispanic americans in the 21st century', *Proceedings of the National Academy of Sciences* **112**(49), 15078–15083.

Gelman, A. & Auerbach, J. (2016), 'Age-aggregation bias in mortality trends', *Proceedings of the National Academy of Sciences* **113**(7), E816–E817.

Kotonya, N., De Cristofaro, P. & De Cristofaro, E. (2018), Of wines and reviews: measuring and modeling the vivino wine social network, *in* '2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)', IEEE, pp. 387–392.

McElreath, R. (2020), *Statistical rethinking: A Bayesian course with examples in R and Stan*, CRC press.

Stigler, S. M. (2006), 'Isaac newton as a probabilist', *Statistical Science* pp. 400–403.

Thorsrud, L. A. (2020), 'Words are the new numbers: A newsy coincident index of the business cycle', *Journal of Business & Economic Statistics* **38**(2), 393–409.

Chapter 2: Statistical Learning and Inference: Basic concepts

Jangho Yang

v1.1

Contents

1 Data, measurement and descriptive data analysis

1.1 Examples of data

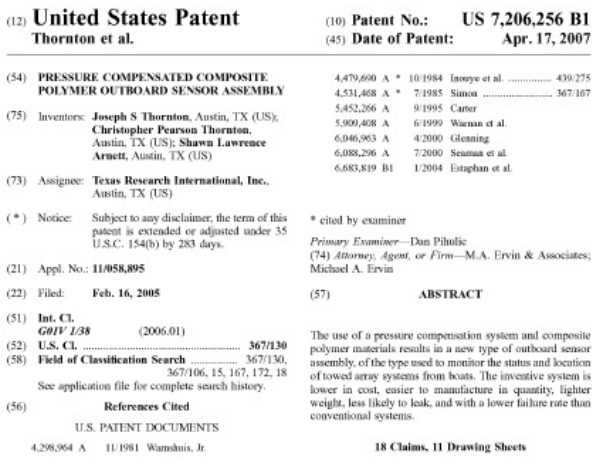
Data (plural) are anything that contains (partial) information about the state of affairs we are interested in. They can be qualitative or quantitative, or a combination of both. See Figures 1, 2, and 3 for examples of different types of data. Figure 1 is an example of survey data where individual respondents provide information about their sex, age, education, income, and so forth. This type of data is used to understand individual/group differences in questions of interest, e.g. voting pattern, perception of political issues, or consumer preference. Figure 2 is an example of corporate financial statements, which are widely used in financial analysis, while Figure 3 is an example of US patent data which can be used to understand the evolution of technological ecosystems.

	Sex	Age	Education	Religion	Income	Ideology
1	female	54	college graduate	protestant	75,000- 99,999	conservative
2	female	27	post-graduate	other	100,000- 149,999	moderate
3	male	56	post-graduate	roman catholic	150,000+	moderate
4	male	64	none	roman catholic	less than 10,000	conservative
5	male	50	college graduate	protestant	75,000- 99,999	conservative
6	male	80	high school graduate	roman catholic	less than 10,000	liberal
7	male	38	college graduate	roman catholic	50,000- 74,999	moderate
8	male	32	post-graduate	roman catholic	20,000- 29,999	liberal
9	male	47	college graduate	protestant	150,000+	conservative
10	female	99	some college	protestant	75,000- 99,999	moderate

Figure 1: Survey data example.

Name	Country	Year	Sector	Sale
American Airlines	USA	1992	4512	14396
American Airlines	USA	2017	4512	42207
Pharmacia	USA	1992	2834	7763
Pharmacia	USA	2002	2834	13993
Canadian Imperial Bank	CAN	1993	6020	10825
Canadian Imperial Bank	CAN	2017	6020	20795
Alberta Energy	CAN	1992	1311	569
Alberta Energy	CAN	2001	1311	6312
Beijing Media	CHN	2003	2700	129
Beijing Media	CHN	2017	2700	61
China Petro & Chem	CHN	2002	5500	37829
China Petro & Chem	CHN	2018	5500	434995
Tempus Holdings	CYM	2010	3845	27
Tempus Holdings	CYM	2017	3845	107
Forbes Ventures	CYM	2007	1000	0
Forbes Ventures	CYM	2013	1000	0

Figure 2: Financial statement data example.



(12) **United States Patent**
Thornton et al.

(10) **Patent No.: US 7,206,256 B1**
(45) **Date of Patent: Apr. 17, 2007**

(54) **PRESSURE COMPENSATED COMPOSITE POLYMER OUTBOARD SENSOR ASSEMBLY**

(75) Inventors: **Joseph S Thornton**, Austin, TX (US); **Christopher Pearson Thornton**, Austin, TX (US); **Shawn Lawrence Arnett**, Austin, TX (US)

(73) Assignee: **Texas Research International, Inc.**, Austin, TX (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 283 days.

(21) Appl. No.: **11/058,895**

(22) Filed: **Feb. 16, 2005**

(51) **Int. Cl.**
G01V 1/38 (2006.01)

(52) **U.S. Cl.** **367/130**

(58) **Field of Classification Search** 367/130, 367/106, 15, 167, 172, 18
See application file for complete search history.

(56) **References Cited**
U.S. PATENT DOCUMENTS

4,298,964 A 11/1981 Warmshuis, Jr.
4,479,690 A * 10/1984 Inouye et al. 439/275
4,531,468 A * 7/1985 Simon 367/167
5,452,266 A 9/1995 Carter
5,909,408 A 6/1999 Warman et al.
6,046,963 A 4/2000 Gleaning
6,088,296 A 7/2000 Seaman et al.
6,683,819 B1 1/2004 Estaphan et al.

* cited by examiner

Primary Examiner—Dan Pihulic
(74) *Attorney, Agent, or Firm*—M.A. Ervin & Associates; Michael A. Ervin

(57) **ABSTRACT**

The use of a pressure compensation system and composite polymer materials results in a new type of outboard sensor assembly, of the type used to monitor the status and location of towed array systems from boats. The inventive system is lower in cost, easier to manufacture in quantity, lighter weight, less likely to leak, and with a lower failure rate than conventional systems.

18 Claims, 11 Drawing Sheets

Figure 3: Patent data example.

1.2 Noisy data and data cleaning

Sometimes, data reveal a great deal about the question of interest, e.g. detailed financial statements of a corporate or security camera footage of criminal activities. But, in many cases, data alone cannot give a clear picture of the object we want to study. Figure 4 showcases the well-behaved data versus noisy data.

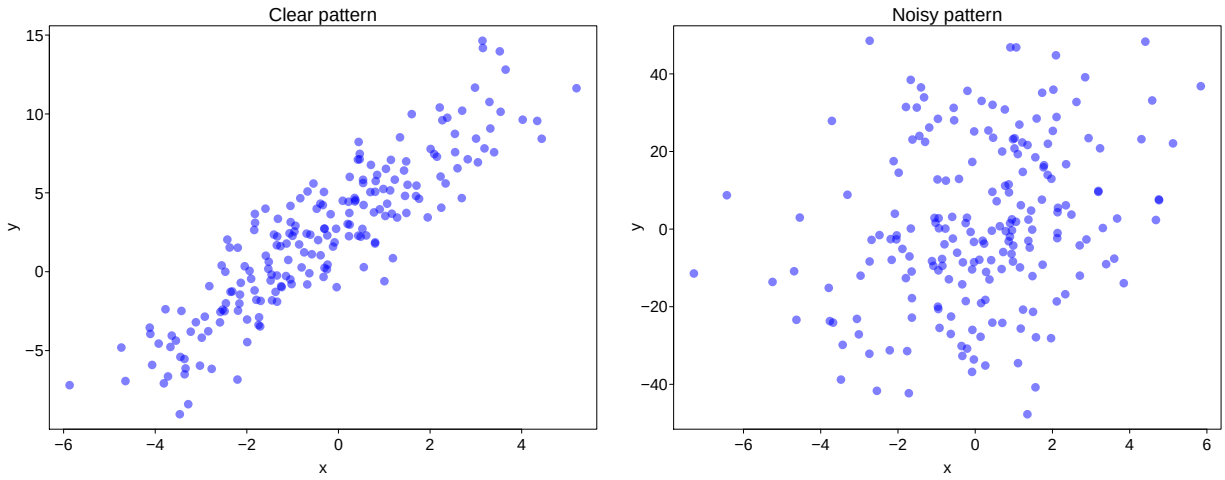


Figure 4: Clear data vs. noisy data.

When data are noisy and do not seem to provide any clear information about the object, it could be due to the fact that the object itself is the result of some complex underlying process. For example, the data on IQ and wealth do not show a clear positive relationship (Zagorsky 2007), which makes sense since there are so many factors other than IQ that determine one's lifetime wealth. Often the case, however, data appear to be noisy due to some nonsensical values that might've come from some human errors in data collecting and reporting process. In this case, some efforts need to be spent *cleaning data* to get them ready for analysis. The problem is that there is no general consensus as to how to clean data. Data cleaning is a grey area in data science from which many research gets awry. For example, Rotemberg & White (2017) notice that many studies based on establishment-level data such as the U.S. Census of Manufactures use somewhat heavily cleaned data (removing many extreme outliers) whose results are not robust under lesser data manipulation (without heavy winsorizing). Figure 5 shows brief summary statistics comparing the original and winsorized data, which reveals substantial differences in standard deviation and other quantile information both in gross output and value-added.

Panel A: Gross Output

	Captured Data			Census-Cleaned Data		
	Outcome			Outcome		
Year	St. Dev	90/10	75/25	St. Dev	90/10	75/25
2002	0.889	1.337	0.577	0.401	0.783	0.331
2007	0.955	1.716	0.902	0.442	0.87	0.356
2012	1.089	1.888	1.031	0.421	0.831	0.346

Notes: The TFPR calculation follow Bils, Klenow and Ruane (2017).

Panel B: Value Added

	Captured Data			Census-Cleaned Data		
	Outcome			Outcome		
Year	St. Dev	90/10	75/25	St. Dev	90/10	75/25
2002	0.981	1.779	0.895	0.575	1.238	0.554
2007	1.1	2.227	1.172	0.616	1.338	0.597
2012	1.256	2.487	1.291	0.626	1.304	0.58

Notes: The TFPR calculation follow Bils, Klenow and Ruane (2017).

Figure 5: Table from Rotemberg & White (2017) showing the difference between the original and cleaned firm-level data.

Then, do we at least have some rules of thumb in cleaning data? Yes, the most important rule is that we need to be transparent and honest about the data cleaning process. Everything needs to be documented with no ambiguity so that others can easily replicate the same results with the same data. An equally important rule is that we need to try hard to keep as much data as possible unless there is conspicuous inconsistency in the data such as changes of data collection/reporting rules, need for the construction of balanced panel, or highly unusual suspicious patterns in the data that seem to come from arbitrary manipulation. Finally, it is important to check whether the results substantially change with different degrees of data cleaning and report how sensitive the results are.

1.3 Descriptive data analysis

Another important rule is to visualize the data as much as we can. By doing so, we can get a sense of whether data are noisy and perhaps why. This initial data summary/visualization is called *descriptive data analysis* and includes summary statistics and overall distributions/time trends of variables of interest. If data has more than one variable, we can also check the pair-wise relationship between variables to see if there are any interesting relations among them. Figure 6 shows an example of the descriptive data analysis for time series (left) and cross-sectional data (right). The left panel is a time-series of 180 products included in the construction of the US consumer price index. A quick visualization reveals that the pattern of price change is highly heterogeneous across products. The right panel is the cross-sectional distributions of the firm-level labour productivity for 4 different industries in France, showing that the productivity in the financial sector is substantially more dispersed compared to other sectors.

Descriptive data analysis sometimes leads to very insightful research. Often the case, high-impact research is based only on descriptive analysis. A notable example is Piketty & Saez (2003, 2006) who collected large administrative datasets to study income/wealth inequality in the world.[1] Figure 7 shows an example of descriptive data analysis using this income database.

[1]Please refer to `https://wid.world` for the latest data on inequality patterns in the major countries.

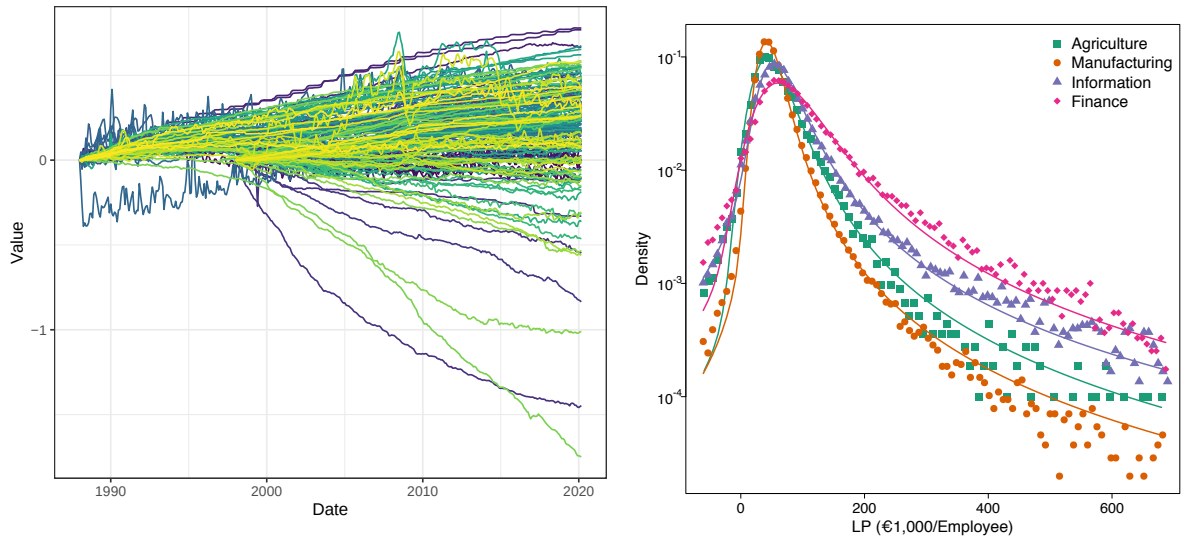


Figure 6: Exploratory data analysis for time series (left) and cross-sectional data (right).

2 Hypothesis and model

2.1 All models are wrong

A *hypothesis* is a claim about a state of nature, and a *statistical hypothesis* is a claim about a state of nature that can be tested with data. To validate the statistical hypothesis, we need *statistical models* that refer to mathematically simplified procedures to approximate some aspects of a state of affairs. It is important to keep in mind that *all models are wrong* since they approximate only a partial aspect of the object. They are useful, however, in the sense that they can help us to understand some complex nature of our hypothesis that is unknown to us prior to modeling.

2.2 Example of statistical hypothesis

Power-law hypothesis

Let's take some examples to get a better sense of how data, hypothesis and model are related to one another. Suppose we are interested in citations in scientific research and want to understand its unequal structure where only a small number of papers get to be cited while the majority are forgotten and get never cited. Here, our statistical hypothesis is that the citation network is dominated by a small number of superstar papers. Since this is a statistical hypothesis, we need data to test it. In this case, the data we are looking for are citation records of scientific publications. How about a statistical model? There are multiple candidates each of which could highlight some aspects of our hypothesis. In this example, let's use a *power-law distribution* as our statistical model. A power-law distribution takes the form of $p(x) = kx^{-\alpha}$ where α is the power-law exponent and k is a constant. To understand what this distribution looks like, suppose $k = 1$ and $\alpha = 2$, then we have $p(x) = x^{-2} = 1/\sqrt{x}$, meaning that when x increases substantially, its probability decreases only by its square root. For example, the probability of $x_1 = 9,000,000$ (a paper being cited 9 million times) is only hundreds times unlikely than $x_2 = 900$ even though x_1 is 10,000 times greater than x_2. This implies that the occurrences of some extreme phenomena are frequent and thus the distribution has a heavy tail. If this power law distribution fits our data well, it means that superstar papers are frequent so that a small number of papers dominate the citation

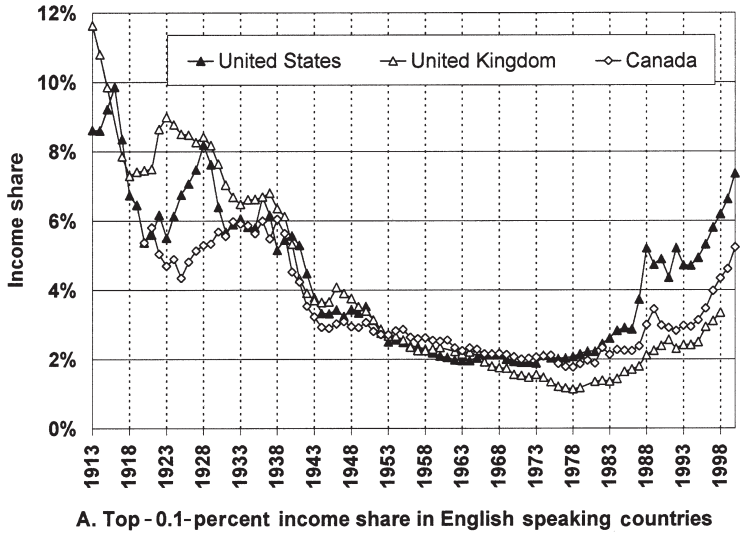


Figure 7: Figure from Piketty & Saez (2006) showing the income share of the top 0.1% income earners in the US, UK, and Canada.

network in scientific research. Figure 8 shows a power-law PDF (left panel) and two example distributions of citations from Redner (1998) (right panel). Note that the power-law distribution exhibits a straight line on a log-log scale. The power-law distribution does seem to fit the citation distribution relatively well as the tail part of the distribution is roughly linear, suggesting that the citation network is indeed dominated by a small number of papers.

Exercise and calories burning

Let's take another simple example. Suppose we are interested in how exercise and calorie burning are correlated. The basic hypothesis is that the more we exercise, the more calories we burn. This hypothesis seems quite benign and we just need data on calorie consumption and the duration of the exercise. However, when we get to the statistical model part, it gets slightly complicated since there are a number of ways to model this relation. Do we want to use a linear relation, like a straight line, suggesting that there are proportionally more calories burnt as we exercise? What if there is some non-linear effect so that the calories do not burn as quickly as before when we already exercised too much? How do we model individual differences in body metabolism? Do we want to use a simple Gaussian error or some other error models? How do we choose the best model among these many possibilities? To answer these questions, we need to discuss estimation and model validation/comparison.

3 Estimation, model validation/comparison, and bias-variance trade-off

3.1 Basic concepts

Let's go back to a linear relationship between two variables. Suppose we are interested in a very simple linear specification only with the intercept and the slope, which we call *parameters* of a model. Now, how can we determine the actual value of these parameters? Equally

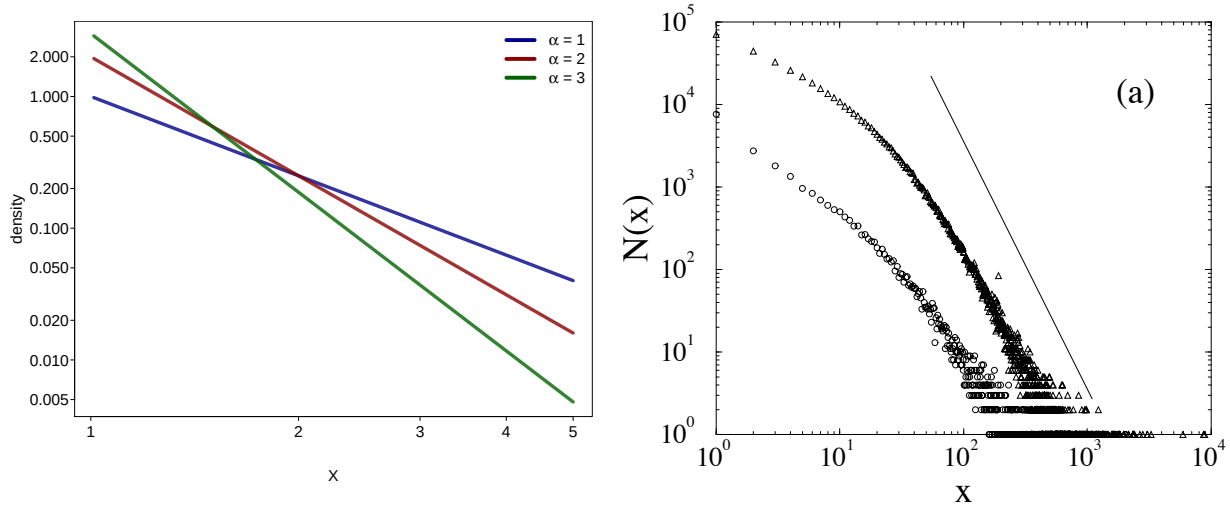


Figure 8: Power law probability density function (left) and the distribution of the citations of 24,296 papers n Physical Review D from Redner (1998). Both are on a log-log scale. Citation distribution from the 783,339 papers in the Institute for Scientific Information (triangle) and the 24,296 papers in the Physical Review D, vols. 11-50 (circle). A straight line of slope 3 is also shown for visual reference in the left plane.

importantly, to which degree of accuracy can we determine such values? Determining the value of model parameters is called an *estimation* problem while checking how accurately a model explains/predicts the data is called a *model validation* problem. If there are multiple models, quantifying the relative accuracy of these models and thus comparing their performance is called a *model comparison* problem. As we will see throughout the course, there is a wide range of procedures for parameter estimation, model validation, and model comparisons.

3.2 Example

To get a concrete example, suppose we have the data on X and Y as shown in Figure 9.

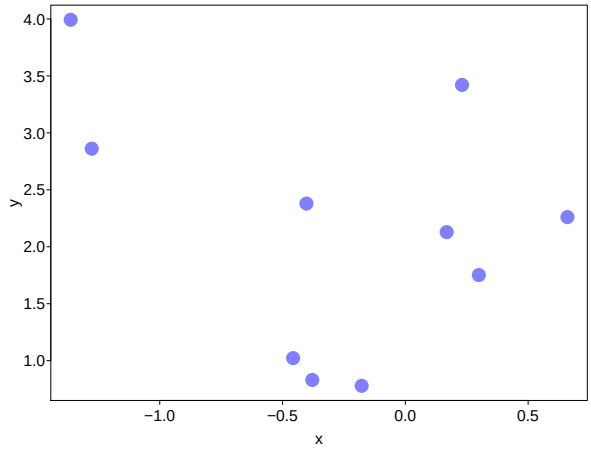


Figure 9: Fake data of x and y

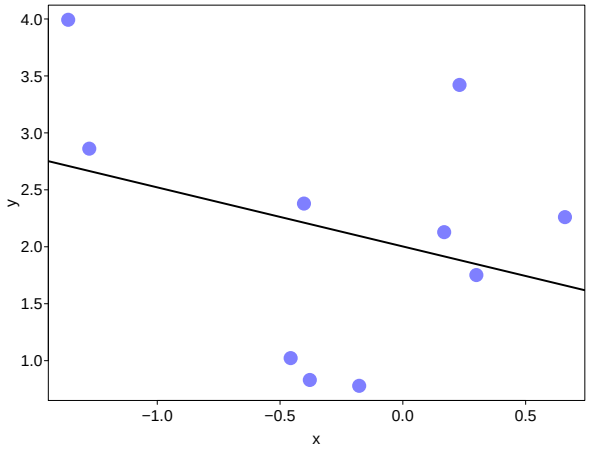


Figure 10: Random regression line.

How many candidate models can we think of for these scatter plots? The simplest possible one is a simple linear specification such that $y = \alpha + \beta x + \varepsilon$ where α is the intercept β is the slope,

and ε is a Normal error term, $\varepsilon \sim \mathrm{N}(0, \sigma^2)$ (which determines the predicted dispersion of data from the linear line.). After we estimate these three parameters (which will be discussed in detail later in this course), we can draw a predicted line over the data as shown in Figure 10. It looks ok but not too satisfactory. How about we make our model a bit more complex by adding more terms as follows?

$$
\begin{aligned}
y &= \alpha + \beta x + \beta_2 x^2 + \varepsilon \\
y &= \alpha + \beta x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon \\
y &= \alpha + \beta x + \beta_2 x^2 + \beta_3 x^3 + \beta_3 x^4 + \varepsilon \\
&\vdots
\end{aligned}
$$

Each of the above equation has increasingly more degrees of freedom with an additional predictor. We visualize the predicted lines in Figure 11. The left plane shows the linear models with 1-3 degrees of freedom and the right plan shows the models with 4-6 degrees of freedom.

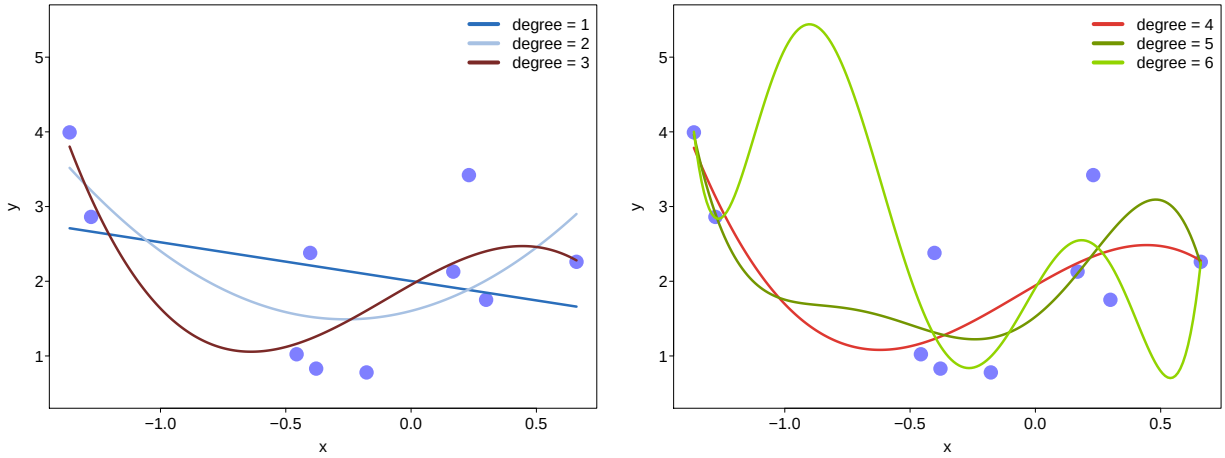


Figure 11: Regression lines with varying polynomial degrees.

Overfitting and bias-variance tradeoff

How can we decide which model among these is more consistent with data? We can see that something quite interesting going on here. The more complex the model is (the higher polynomial degrees), the better it fits the data. For example, the model with 6 degrees of freedom have all the data points very close to the predicted line. So, the model in the bottom right appears to perform the best. If this is the case, can we say that the most complex model is always preferable because it always gives the best fit?

Here we encounter the issue of *overfitting*. To understand the pitfall of overfitting, we need to distinguish *the in-sample* validation from the *out-of-sample* validation. When we determine the model performance only using the data we use for model estimation, this is the in-sample validation. In contrast, when we determine the model performance using the "future" data that is not included in the current dataset, we call it out-of-sample validation. Complex models

are very good at in-sample performance but are normally bad at out-of-sample performance, simply because complex models are too tight and inflexible to accommodate the noisy reality of the world. This is called *bias-variance trade-off*. That is, complex models have low bias and explain the in-sample data accurately. But they are extremely sensitive to change in the data set and thus have high variance in explaining future data (or another set of in-sample data).

To better understand this bias-variance trade-off, let's generate a few more data points using the exact same data generating process as above. We consider these data points to be "future" data to see how well each of the models predicts these data points. We compare the polynomial model with degrees of 2 and 6 in Figure 12.

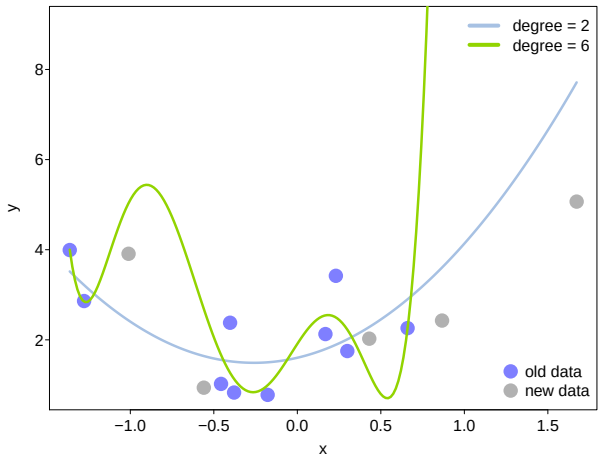


Figure 12: Regression lines with polynomial degrees of 2 and 6, predicting new data points.

Obviously, the complex model with a degree of 6 performs poorly and its predicted line is far away from the new data points. This is generally the case for most complex models. In contrast, a simple model with a degree of 2 does a good job of predicting the new data points. Many model comparison techniques have been proposed to account for both or in-sample and out-of-sample validation, e.g. AIC, DIC, Cross-validation, some of which will be covered in this course.

References

Piketty, T. & Saez, E. (2003), 'Income inequality in the united states, 1913–1998', *The Quarterly journal of economics* **118**(1), 1–41.

Piketty, T. & Saez, E. (2006), 'The evolution of top incomes: a historical and international perspective', *American economic review* **96**(2), 200–205.

Redner, S. (1998), 'How popular is your paper? an empirical study of the citation distribution', *The European Physical Journal B-Condensed Matter and Complex Systems* **4**(2), 131–134.

Rotemberg, M. & White, T. K. (2017), Measuring cross-country differences in misallocation, *in* 'North East Universities Consortium annual conference, Tufts University, Medford, MA, November', pp. 4–5.

Zagorsky, J. L. (2007), 'Do you have to be smart to be rich? the impact of iq on wealth, income and financial distress', *Intelligence* **35**(5), 489–501.

Chapter 3: Statistical Learning and Inference: Competing methods

Jangho Yang

v1.0

Contents

As we discussed in Topic 1, the probability is not a monolithic concept. Quite the contrary, there are opposing views of the probability that lead to different procedures for statistical inference. The most conventional one is what we call the classical/frequentist approach, which will be discussed in great detail throughout the course. The lesser-known but equally important one is the Bayesian approach and will be discussed in tandem with the classical approach. We will also introduce the maximum likelihood approach since this is the most frequently used estimation method.

1 Overview of the classical approach

1.1 True data-generating process, repeated experiments, and estimator

The classical/frequentist approach refers to a broad set of statistical procedures based on sampling theory. Classical statistical methods presuppose a true *data-generating process* or a true *model*. Understanding this "true" nature of the state can be achieved by an infinite sequence of trials/experiments. That is, if we can repeat the experiments infinitely many times (sampling from the *population*), we can theoretically obtain the true nature of the state. In reality, we can't really repeat the same experiment forever and thus the data we have is always finite. Then, how can we approximate the true data-generating process with finite trials?

Let's unpack this using a simple statistical exercise. Suppose that we want to understand the data-generating process behind the sequence of heads (H) and tails (T) generated by tossing a coin multiple times, $\{H, T, T, H, \dots\}$. Here, the key part of the data-generating process is a probability p of a head (which we call *parameter*), and a probability $1 - p$ of a tail by symmetry. Given the probability p, we can easily calculate the probability of any sequence of H and T. Suppose we have n_H number of heads and n_T number of tails in n total number of trials, $n = n_H + n_T$. Assuming independence of each trial, the probability of seeing n_H and n_T is simply the binomial distribution

$$\binom{n}{n_H} p^{n_H} (1-p)^{n-n_H} \tag{1}$$

Given this simple data generating process of coin tossing, what classical statisticians want is to recover the true value of p that generates the sequence of heads and tails we are observing. Again, we cannot do that simply because we do not have an infinite sequence of data. What we can do instead is to find some function, which we call $\hat{p}$, that is assumed to have the closest relationship with the true parameter p. Classical statisticians call this proxy for a true parameter an *estimator*.

Desirable properties of an estimator

Then, what is the right estimator in our coin-tossing example? We can come up with many different estimators that we hope to have some relationship with the true parameter. Classical statisticians have a set of criteria to evaluate the quality of an estimator. Three key criteria are *unbiasedness*, *consistency*, and *efficiency*. Let's get some intuition for each of these concepts.

Unbiasedness: The expected value of an estimator is equal to the true parameter: $\mathrm{E}[\hat{p}] = p$

Consistency: An estimator converges to the true value as the size of data increases: $\hat{p}_n \to p$ in probability as $n \to \infty$ where $\hat{p}_n$ is an an estimator of p with the sample size n.

Efficiency: The variance of an estimator is as small as possible, e.g. $\min \mathrm{E}[\hat{p} - p]^2$.

A great deal about classical statistics is to mathematically prove whether a certain estimator has these particular properties. The same goes for our coin-tossing example, a simple sequence of Bernoulli trials. One natural choice of a function (of data) that has a close relationship with the "true" probability of heads is $\hat{p} = n_H/n$, a simple proportion of the heads in the data. We can mathematically prove that this particular estimator is actually an unbiased, consistent, and efficient estimator. We will get back to this point later in the course.

1.2 Sampling distribution and confidence interval

Setting aside these nitty-gritty properties of an estimator, let's stop and think about how much we can trust this estimator. Again, any estimator is a function of observed data, like n_H and n in our coin-tossing example. Suppose one extreme case where we have only one observation (only one coin has been tossed) and we observe a head. Using the best estimator we discussed above, we come to a conclusion that $\hat{p} = n_H/n = 1$, that is the coin is completely biased against a head. Does this really make sense? No, and it is highly likely that our estimator $\hat{p}$ might be considerably different from the true parameter p. This simple exercise hints at one important point: we would be less confident about our estimator when the sample size is too small. Then, how do we quantify our confidence about an estimator?

Repeated trials and sampling distribution

Here, we need some thought experiments. Suppose we are able to replicate the exact same experiment many times and record the result of the quantity of our interest. We then take the distribution of these multiple results from the repeated trials. If the sample size is small, this distribution will be more widely dispersed, meaning that there is a high degree of sampling error. In contrast, if the sample size is large and closer to the entire population, the distribution will be less dispersed with a very narrow range, meaning that we have little sampling error and thus the repeated experiments give a similar answer all the time. The classical statisticians call this distribution from the repeated trials a *sampling distribution*. It is a theoretical (or imaginary) distribution of whatever quantity of finite data at hand (statistic), e.g. its mean or median, constructed by generating an infinite number of such quantity from the true data-generating process. From this sampling distribution, the classical statisticians construct a *confidence interval*, the tolerance range of the estimator given the pre-determined confidence level.

Let's unpack this using a coin-tossing example. In this example, the quantity of data of our interest is the proportion of heads, $\hat{p}$, in a sequence of heads and tails. Then, the sampling distribution of $\hat{p}$ can be constructed by repeatedly calculating the proportion of heads from some imaginary data of the same length generated from the true data-generating process of a coin tossing. If we can repeat this process infinitely many times, we can get a sampling distribution. Let's simulate some of this process. Even though we will not be able to simulate it infinitely, we can get some idea of how the sampling distribution of $\hat{p}$ looks like. Figure 1 shows several different numbers of trials of tossing a coin 100 times and recording the proportion of landing heads. The number of trials varies from 100 to 100,000. As we increase the number of trials, the distribution becomes more orderly with a clear shape. The sampling distribution is derived by repeating this trial infinitely.

Some estimators statistical problems have a *well-defined* sampling distribution for estimators in the sense that the distribution is analytically defined with a particular functional form. This is the

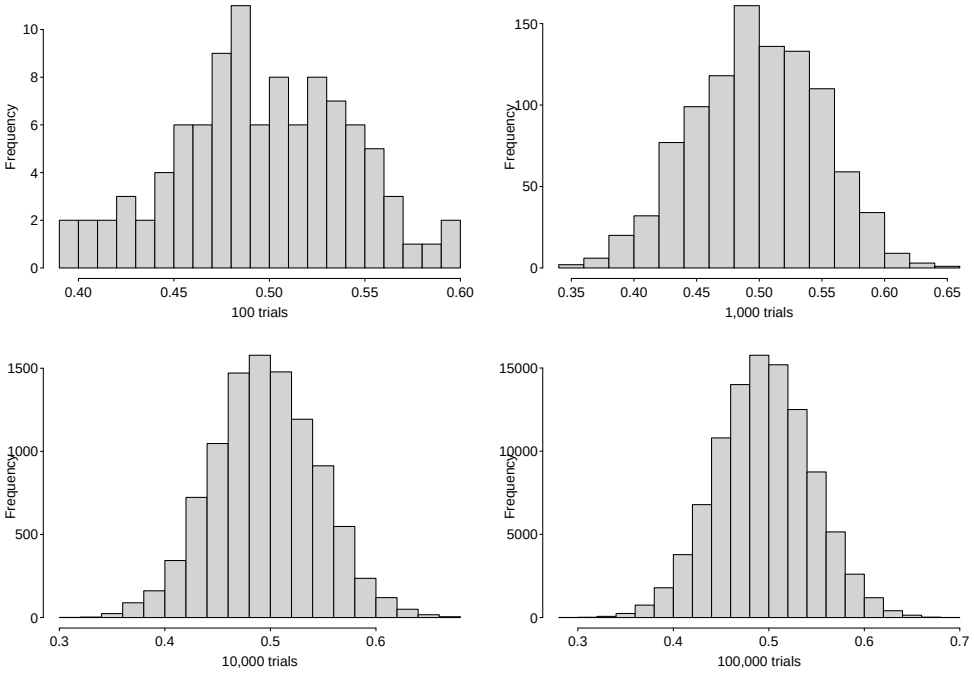


Figure 1: Simulations of tossing a coin 100 times with varying number of trials.

case with a coin-tossing example. With a bit of math, we can show that the sampling distribution of $\hat{p}$ can be approximated by a normal distribution. This also can be shown by fitting a normal distribution to the simulated sampling distribution with a large number of trials. Figure 2 shows a normal distribution over the histogram of $\hat{p}$ from 100,000 trials. As expected, a normal distribution gives a good fit.

Determining confidence intervals

With a particular sampling distribution at hand, we are almost done with calculating a confidence interval. We then need to specify what percentage of area under the distribution we want to show. The percentage of area under the curve is called *confidence level*. By definition, the maximum confidence level is 100%, meaning that we want to use the entire range of the distribution as our confidence interval. Practitioners in classical statistics often use 99%, 95% and 90% as rule of thumb intervals. Let's go back to the coin-tossing example. With a normal distribution as the sampling distribution of $\hat{p}$ and with a 95% confidence level, the confidence interval of the probability of a head is distributed symmetrically around $\hat{p}$ as follows:

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}} \tag{2}$$

where n is the sample size. Again, it is important to note that this interval is derived from the assumption that the sampling distribution follows a normal: $\hat{p} \sim \mathrm{N}(p, \sqrt{p(1-p)/n}^2)$. Note that the variance of the binomial distribution is $p(1-p)/n$.[1] The derivation of this interval needs standardization of the random variable so that $\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \sim \mathrm{N}(0,1)$. Let's do some sanity check to see if this confidence interval makes sense. First, as we discussed, it is indeed a function of

[1]See Topic 1 for an overview of the binomial distribution.

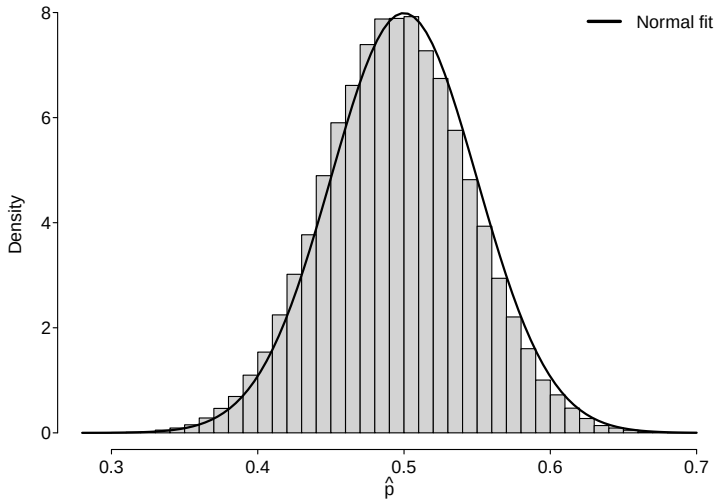


Figure 2: Simulations of tossing a coin 100 times with 100,000 trials. A normal fit is overlaid.

the estimator $\hat{p}$ and the sample size, n. Second, as the sample size increases (decreases), the confidence interval becomes smaller (larger).[2] This means that when we have more data (larger sample size), we can determine the range of the interval more narrowly.

Note that the confidence interval is not the interval of potential values of the quantity of our interest. That is, a 95% confidence interval of p in our coin-tossing example is not a 95% range of all possible values of p. This interpretation is simply not allowed in classical statistics since the whole purpose of inference is to find the "true" value of p, a constant, with higher precision as possible. Rather, a confidence interval should be interpreted as the range within which the "true" value of our statistic lies when we repeat the same experiment indefinitely. Therefore, a 95% confidence interval of p is the range where the true value of p lies with a 95% chance when we repeat the same coin-tossing experiment many times.

1.3 Hypothesis testing

Hypothesis testing and the null hypothesis

Classical statisticians go one step further and make use of a confidence interval in evaluating particular statistical hypotheses, which is often called *hypothesis testing*. The key idea is to derive a confidence interval according to a certain hypothesis, often called the *null hypothesis*, and check the value of the estimated parameter (derived from the observed data) and ask whether this value lies within the corresponding confidence interval. For example, suppose we want to test a statistical hypothesis that the coin is fair while the observed proportion of heads is 0.6 after 5 trials. Given the data and the predetermined confidence level, we can check whether this hypothesis that the coin is fair $p = 0.5$ is consistent with the observed proportion 0.6.

P-value

[2]One caveat with this particular confidence interval for a repeated Bernoulli process is that the sample size n cannot be too small and $\hat{p}$ cannot be too close to 0 or 1 due to the required assumptions for a Normal approximation to Binomial distribution.

A standard procedure to check whether an observed value of our data is within the given confidence interval is done by calculating its *p-value* and comparing it to a pre-defined significance level. This p-value picks up a tail probability of the observed data given the sampling distribution according to our null hypothesis. In the coin-tossing example, our null hypothesis is "the coin is fair" and the test statistic of our interest is the proportion of heads p, and, most importantly, we can construct the sampling distribution of this test statistic by repeating the (fair) coin-tossing trials many times. Using the normal distribution as our sampling distribution, we have $\hat{p} \sim \text{N}(p, \sqrt{p(1-p)/n}^2)$, where $p = 0.5$ (null hypothesis that the coin is fair) and $n = 5$ (the number of trials). We visualize the sampling distribution in Figure 3.

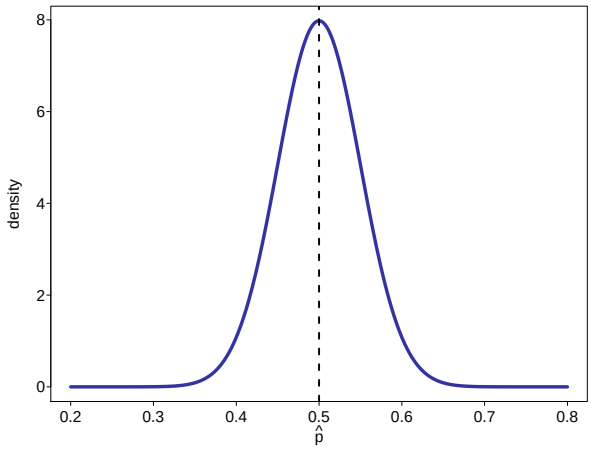


Figure 3: A normal sampling distribution of the null hypothesis that $p = 0.5$ with $n = 5$.

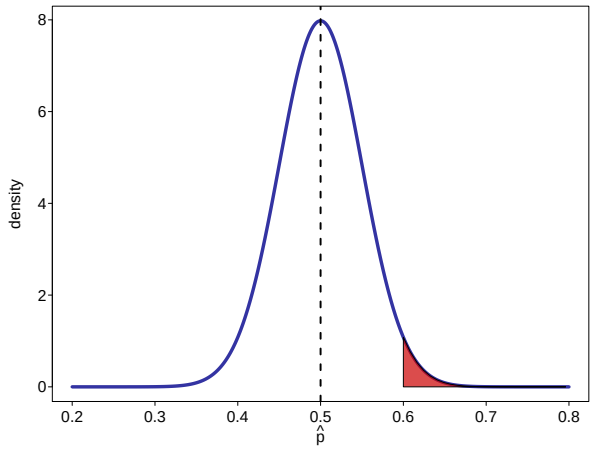


Figure 4: A normal sampling distribution with p-value.

Again, suppose the observed proportion of heads in our data is 0.6. Here, the p-value is the probability of test results being equal to or greater than 0.6. Formally, the *p-value* is defined as the probability that the test results are as extreme or more extreme than the observed data under the null hypothesis sampling distribution

$$P(T(y^{rep})) > T(y|H), \tag{3}$$

where T is a test statistic of our interest, e.g. mean, median or proportion of our data, y^{rep} is the test results obtained by repeating the same trials many times, y is the observed data, and H is the null hypothesis. Visually, the p-value is the area under the distribution whose test results are equal to or greater than the observed data. In our coin-tossing case, the p-value is the area under the normal distribution greater than 0.6. Figure 4 shows the p-value for our coin-tossing trial.

To complete the hypothesis testing, we just need to compare the calculated p-value with the pre-determined significance level (1-confidence level/100), often called α. If the p-value is smaller than the significance level, we "reject" the null hypothesis. For our coin-tossing exercise, we can calculate the p-value of $\hat{p} = 0.6$ by taking the integral over the area equal to or greater than 0.6, which 0.023. Suppose our significance level is 0.05. Since the p-value is smaller than the given α level, we reject the hypothesis that the coin is fair. This means that $\hat{p} = 0.6$ (the estimated parameter from the data) is unlikely to result from our sampling distribution by chance given that $p = 0.5$. However, if our α is 0.01, which is a more stringent criterion, we cannot reject our

hypothesis that the coin is fair.

1.4 Remarks on p-value

Having a strict threshold for validating/falsifying a certain hypothesis might be useful when we have to make a decision based on some statistical testing, e.g. policy recommendation. Unfortunately, almost all statistical problems do not have a black and white answer and the grey area is often too wide to draw any definite conclusions from. There are many reasons for this. Data could be too noisy and complex and/or there is a wide range of different models for data. When dealing with complex data and models, researchers face myriad decisions to make in every step of statistical analysis, each of which will lead to different results in terms of p-value and the statistical significance of the models.

The problem arises when the researchers modify the data selection process and/or try out different model specifications until nonsignificant results become significant. This is called *p-hacking* or *data fishing*. Unfortunately, there is no fundamental remedy to p-hacking since the dichotomous aspects of the p-value is deeply embedded in the classical hypothesis testing. Recently, the focus of statistical inference has moved from narrowly defined hypothesis testing to a broad set of model validation based on out-of-sample prediction. Even when the hypothesis testing is used, it is often supplemented by additional predictive checking to make sure that the model is not picking up some random noise in the data. We will be discussing some useful model validation/comparison techniques later in the course.

1.5 Examples of the sampling distribution for hypothesis testing

Without a question, a normal distribution is the most widely used sampling distribution for hypothesis testing due to its generality resulting from the central limit theorem. See Topic 1 for an overview of a normal distribution. Here, we introduce two more sampling distributions, each of which will be used in different statistical problems/hypothesis testing.

Chi-square distribution

When we are interested in estimating the variance of the normally-distributed random variable, the sampling distribution of the sample variance is the Chi-square distribution. This distribution is defined as a sum of the squares of k independent standard normal random variables. Let $Z_1, Z_2, \ldots, Z_k$ be a standard normal random variable $\mathrm{N}(0,1)$ and let $X = Z_1^2 + Z_2^2, \ldots, +Z_k^2$, the PDF of X is

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)}\, x^{k/2-1} e^{-x/2} \qquad \text{for } x \geq 0 \tag{4}$$

The expected value and the variance are k and $2k$, respectively. k is known as the degrees of freedom and determines the shape of the distribution. Figure 5 shows the Chi-Square PDF with varying degrees of freedom.

Student-t distribution

When we are interested in estimating the mean of a normally-distributed random variable with unknown variance (especially with small sample size), the sampling distribution of the mean is Student's t-distribution. It has a heavier tail than the normal distribution and thus can reflect a

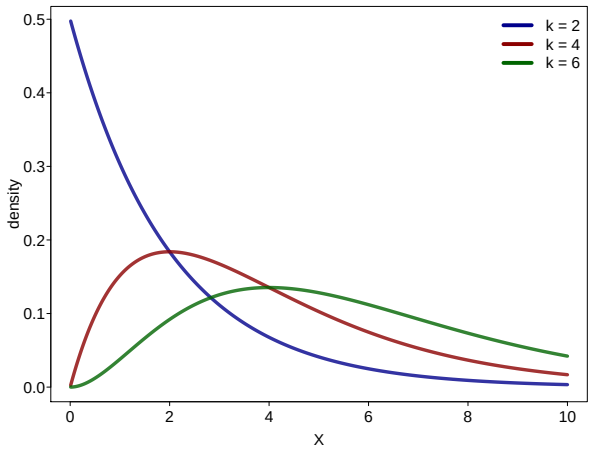


Figure 5: Chi-Square probability density function with varying parameters of the degrees of freedom k.

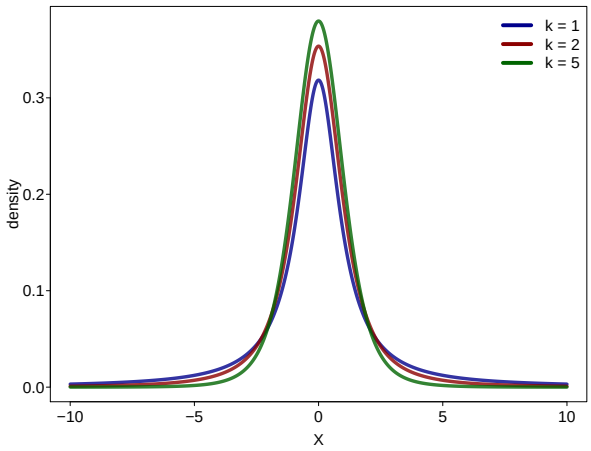


Figure 6: Student's t probability density function with varying parameters of the degrees of freedom k.

higher uncertainty in estimating the mean when the variance is unknown with small sample size. The t-distribution is defined with respect to the (standard) normal distribution and the Chi-square distribution. Let $Z \sim N(0,1)$ and $V \sim \chi^2_k$, then the student-t distribution is

$$\frac{Z}{\sqrt{V/k}} \tag{5}$$

Note that when the degree of freedom is 1, the distribution becomes pathological and the mean and the standard deviation are not defined. When $k > 1$, the mean is zero and the standard deviation is either $k/k-2$ when $k > 2$ or infinity when $1 < k < 2$. Figure 6 shows the Student's t PDF with varying degrees of freedom.

1.6 Example

Figure 7 shows the histograms of annual price change for four different products: integrated microcircuits, personal computer, eggplants, and wastepaper. The mean annual price change for all products (red line) is below zero (black dotted line). However, the price change of the first two products is systematically below zero while that of the other two products fluctuates around zero. How can we statistically determine that the mean price change of each product is different from zero?

A standard approach from the classical point of view is

(i) Define the test statistics

(ii) Derive the sampling distribution of the test statistics

(iii) Set up a hypothesis regarding a specific value of the test statistics

(iv) Decide the confidence level (α) to reject/accept the hypothesis

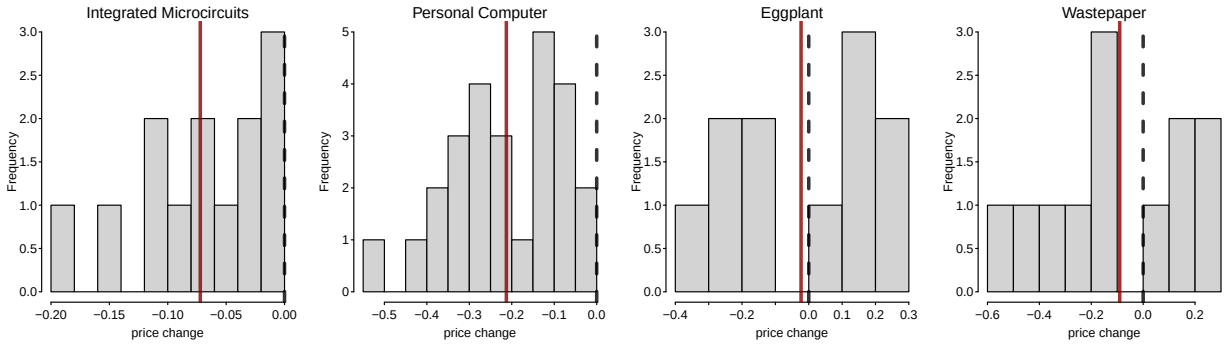


Figure 7: Histogram of product price changes for four products from US Consumer Price Index (CPI). The red solid line represents the mean price change and the black dotted line represents 0.

(v) Calculate the p-value using the data and compare the p-value with the pre-determined confidence level

For this problem,

(i) the test statistics is the mean price change, which we will denote by $\bar{X}$. We will be assuming that $\bar{X}$ is normally distributed with mean μ and the standard deviation $\sqrt{\sigma/n}$ (due to the central limit theorem). Both μ and σ are unknown and need to be estimated.

(ii) In this setting, we can show that the sampling distribution for the estimator of μ, $\hat{\mu}$, follows the student's t distribution.[3]

(iii) The null hypothesis is that the mean price change is zero: $\mu = 0$.

(iv) Suppose the confidence level is 5% ($\alpha = 0.05$).

(v) Using student's t distribution as a sampling distribution and with the null hypothesis $\mu = 0$, the p-values for each of 4 sets of the observed data of price changes are 0.0003846, 0.00000001787, 0.7348, and 0.2615, respectively. When comparing these p-values with $\alpha = 0.05$, the first two p-values are smaller and the latter two p-values are bigger.

Therefore, we reject the hypothesis that $\mu = 0$ for the first two products (integrated microcircuits and personal computer) but can't reject it for the latter two products (eggplants and wastepaper).

2 Overview of maximum likelihood approach

2.1 Definition of MLE

Statistical problems as counting

Sometimes, it is easier to think of statistical problems as counting up all the ways data can happen according to given hypotheses. For example, suppose we have a bag of four balls with two different colors of black and white, and we want to guess the number of black balls in a bag after drawing three balls without replacement. Suppose we drew two black balls and one white ball. Since there are 4 balls in total, 5 hypotheses are i) 4 black balls and 0 white ball, ii) 3 black balls and 1 white ball, iii) 2 black balls and 2 white balls, iv) 1 black ball and 3 white balls, and v) 0

[3]See Topic 4 for a more detailed discussion on this in the context of linear regression.

black ball and 4 white balls. For each hypothesis, we can count up all the ways the observed data of two black balls and one white ball can happen as follows:

(i) 4 black balls and 0 white ball: $\binom{4}{2}\binom{0}{1} = 0$

(ii) 3 black balls and 1 white ball: $\binom{3}{2}\binom{1}{1} = \mathbf{3}$

(iii) 2 black balls and 2 white balls: $\binom{2}{2}\binom{2}{1} = 2$

(iv) 1 black ball and 3 white balls: $\binom{1}{2}\binom{3}{1} = 0$

(v) 0 black ball and 4 white balls: $\binom{0}{2}\binom{3}{1} = 0$

We can see that the hypothesis of 3 black balls and 1 white ball has more ways to realize the observed data and thus more probable/plausible. The maximum likelihood estimation (MLE) is nothing but finding the hypothesis that is most likely and thus has the greatest number of ways to realize the observed data.

Formal definition of MLE

Formally, the MLE is a statistical procedure that finds the point estimate of parameters by maximizing a likelihood function associated with them. Let's unpack this one by one. The likelihood is defined as the probability of data given hypotheses, that is, the number of ways the data can realize according to hypotheses:

$$p(y|\theta) \tag{6}$$

where y is data and θ is hypotheses. Note that this likelihood is a function of hypotheses, meaning that the function takes varying hypotheses as input and gives a likelihood value for each hypothesis, which can be understood as a relative probability that can be compared across different hypotheses. However, the likelihood itself is not a proper probability distribution because its sum with respect to θ (the sum of all likelihood values for all hypotheses) is not 1. To highlight its difference from a probability distribution, a standard notation for the likelihood function uses $\mathcal{L}$ as follows:

$$\mathcal{L}(\theta|y) \tag{7}$$

This notation makes it more clear that the likelihood function is a function of hypotheses θ given data y. For each hypothesis, $\mathcal{L}(\theta|y)$ is calculated by taking the joint probability of data, which can be expressed as the product of all probabilities of data given the hypothesis: $\prod_i^n p(y_n|\theta)$. Among all the likelihood values for all hypotheses in the parameter space Θ, the MLE finds the hypothesis $\hat{\theta}$ that has the maximum likelihood value[4]

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \widehat{\mathcal{L}}_n(\theta\,|y) \tag{8}$$

2.2 MLE and optimization algorithm

Then, how can we find the maximum likelihood value? In principle, when the likelihood function is differentiable, we can find the first-order condition $L^{'} = 0$ and check the second-order derivative is negative $L^{''} < 0$. However, for the vast majority of statistical problems, taking the first and

[4]Note that $\max f(x)$ represents the maximum value of $f(x)$ while $\arg\max f(x)$ represents the value of x at which the maximum of $f(x)$ is attained.

the second derivative is extremely challenging or such derivatives might not exist. In this case, we need to rely on *numerical optimization* to find the maximum value numerically, not analytically. There is a wide range of optimization algorithms that help us to guess the maximum/minimum of a likelihood function. Going through these algorithms in detail is beyond the scope of this course. Instead, we will directly implement some of these algorithms in the MLE examples we will discuss in this course.

2.3 Normal approximation to MLE estimators

Before we discuss some examples of MLE, let us quickly point out one missing block in our MLE approach, that is the interval estimation in the MLE, e.g. confidence interval. One of the most important properties of the maximum likelihood estimator is that, under certain conditions, the maximum likelihood estimator converges in distribution to a normal distribution. This means that we can easily construct a confidence interval by using the normal distribution as a sampling distribution as we discussed above.[5]

2.4 Example

We now turn our attention to an example of how to implement the MLE in actual statistical problems. Let's revisit our coin-tossing example above where we tried to find the true p (the probability of a head) using the observed proportion of landing heads as the estimator $\hat{p}$. We had the proportion of landing heads = 0.6 with n = 5, meaning 3 heads out of 5. In the MLE framework, each possible value of $\hat{p}$ is a hypothesis to which a certain likelihood is assigned. To get some intuition of how this works, let's manually calculate the likelihood for some plausible values of $\hat{p}$, using the binomial PMF, $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$

(i) Suppose $\hat{p} = 0.4$. Then, having 3 heads out of 5 is $P(X = 3) = \binom{5}{3} 0.4^3 (1 - 0.4)^{5-3} = 0.230$

(ii) Suppose $\hat{p} = 0.45$. Then, $P(X = 3) = \binom{5}{3} 0.45^3 (1 - 0.45)^{5-3} = 0.276$

(iii) Suppose $\hat{p} = 0.5$. Then, $P(X = 3) = \binom{5}{3} 0.5^3 (1 - 0.5)^{5-3} = 0.312$

(iv) Suppose $\hat{p} = 0.55$. Then, $P(X = 3) = \binom{5}{3} 0.55^3 (1 - 0.55)^{5-3} = 0.337$

(v) Suppose $\hat{p} = 0.6$. Then, $P(X = 3) = \binom{5}{3} 0.6^3 (1 - 0.6)^{5-3} = \mathbf{0.346}$

(vi) Suppose $\hat{p} = 0.65$. Then, $P(X = 3) = \binom{5}{3} 0.65^3 (1 - 0.65)^{5-3} = 0.336$

(vii) Suppose $\hat{p} = 0.7$. Then, $P(X = 3) = \binom{5}{3} 0.7^3 (1 - 0.7)^{5-3} = 0.309$

(viii) Suppose $\hat{p} = 0.75$. Then, $P(X = 3) = \binom{5}{3} 0.75^3 (1 - 0.75)^{5-3} = 0.264$

(ix) Suppose $\hat{p} = 0.8$. Then, $P(X = 3) = \binom{5}{3} 0.8^3 (1 - 0.8)^{5-3} = 0.205$

Among the 9 hypotheses listed above, the hypothesis of $\hat{p} = 0.6$ has the highest likelihood, meaning that this is most consistent with the observed data. Figure 8 visually shows the likelihood of more hypotheses: $0 < \hat{p} < 1$, also confirming that $\hat{p} = 0.6$ has the highest likelihood and is most likely.

[5] Finding the sampling distribution of the MLE estimators involves finding variance-covariance matrix, which needs to be calculated by the inverse of the *Information matrix*, the negative of the expected value of the Hessian matrix. A detailed discussion on this requires some matrix algebra and thus will not be covered in the course.

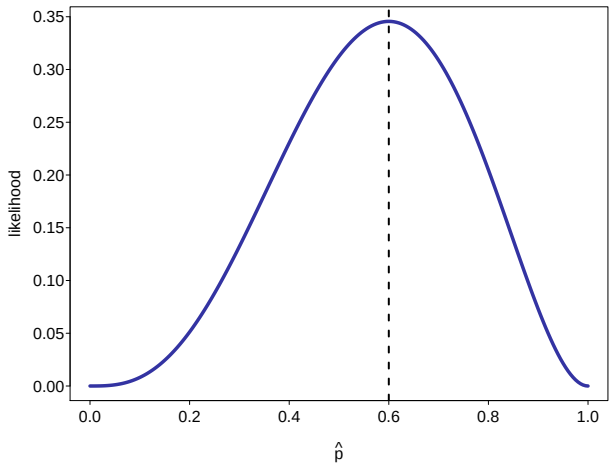


Figure 8: Likelihood function for a coin tossing with n = 5.

This result implies that the MLE estimator is equivalent to the classical estimator for p that uses the proportion of heads, $\hat{p} = n_H/n$. For this simple binomial case, we can mathematically derive the MLE estimator and show that it is equal to n_H/n.

The likelihood function for the binomial trial can be written as follows.

$$\begin{aligned}\widehat{L}_n(p\,|x,n) &= \prod_i^n \binom{n}{x_i} p^{x_i}(1-p)^{n-x_i} \\ &= \left(\prod_i^n \frac{n!}{x_i!(n-x_i)!}\right) p^{\sum_i^n x_i}(1-p)^{n-\sum_i^n x_i}\end{aligned}$$

It is often more convenient to work with the log-likelihood since a summation is more tractable than multiplication from the computational perspective. The log-likelihood function for the binomial trials is written as follows

$$\widehat{l}_n(p\,|x,n) = \sum_i^n x_i \log(p) + \left(n - \sum_i^n x_i\right)\log(1-p) + C \tag{9}$$

where $\widehat{l}_n$ represents a log-likelihood function and $C = \sum_i^n \log(\frac{n!}{x_i!(n-x_i)!})$ is a constant. Therefore, the MLE for the binomial trials boils down to solving the following maximization problem

$$\hat{p} = \arg\max_p \widehat{l}_n(p\,|x,n) \tag{10}$$

The solution can be found by taking the first derivative of this log-likelihood function with respect to p and set it to zero

$$\begin{aligned}\frac{\partial \widehat{l}_n(p\,|x,n)}{\partial p} &= \frac{1}{p}\sum_i^n x_i + \frac{1}{1-p}\left(n - \sum_i^n x_i\right) = 0 \\ \hat{p} &= \frac{\sum_i^n x_i}{n}\end{aligned} \tag{11}$$

The result shows that the MLE estimators for $\hat{p}$ is exactly the same as the classical estimator we discussed above.

3 Overview of Bayesian approach

3.1 Prior information

The likelihood-based approach we discussed above is nothing but counting up all the ways that data can realize according to hypotheses. A Bayesian approach goes one step further and utilizes prior knowledge about the degree of plausibility of some hypotheses. For example, suppose you have never seen any biased coins in your life and you are asked to estimate the probability of heads after seeing 3 heads out of 5 tossings as above. Unlike the simple MLE procedure, you want to use your prior information that you haven't seen any biased coins and put less weights on the probability that p is really close to either 0 or 1. For another example, suppose that there is a well-established medical trial where the effectiveness of a certain drug has been tested. The initial test showed that the drug can be effective 8 out of 10 times. Then, you're asked to try the same drug for different patient groups. How would you utilize the prior test results in your new experiment? Finally, suppose you have a highly complicated model, which you believe provides good insight into the problem you're grappling with. The issue is that the model is too complicated so you can't really estimate its parameters. However, you know that some values of the parameters are not feasible in the real world and can be excluded from the beginning, which will help the model implementation a lot easier. As we will see in this section, a Bayesian approach enables us to make use of the prior information/knowledge/assumption (prior to experiments) in our statistical inference in a systematic manner.

3.2 Bayes' theorem and Bayesian inference

One great advantage of using a Bayesian approach lies in its conceptual simplicity and coherency. This strength comes from the fact that the Bayesian method starts and ends with a very simple probability rule, called Bayes' theorem:

$$P(\theta|y) = P(\theta)P(y|\theta)/P(y), \tag{12}$$

where θ represents hypotheses and y is data. The ultimate goal of the Bayesian statistical analysis is to get $P(\theta|y)$, the posterior probability of hypotheses given data. Its literal meaning is the probability of our guess about the state of affairs (hypotheses) when we have some partial evidence of it (data). When working with the probabilistic form of hypotheses ("parameters as a random variable"), we can get the degree of plausibility of a set of all hypotheses. Notice how this approach considers θ to be a random variable while the classical or frequentist approach considers θ to be a constant (unknown). For example, suppose we have 15 balls with two different colors of black and white in an urn and we want to guess how many black balls are in the urn after drawing 5 balls from it. In this experiment, θ is the number of black balls in the urn and y is the color of 5 balls we have drawn. Here, Bayesians assign a probability to each of all 16 hypotheses from # of black ball = 0 to # of black ball = 15, after observing how many black balls drawn from the urn. Note that each outcome has its own probability and the sum of all probabilities should be 1, meaning that $P(\theta|y)$ is a proper probability distribution. By forming a probability distribution of our hypotheses, we can quantify how plausible each of these hypotheses is given data (or put it differently, how consistent each of these hypotheses is with data).

Then, what is the underlying assumption that makes it possible to form a probability distribution of hypotheses? The fact that we can assign a probability to each of all

possible outcomes implies that there is no single true state of affairs (or "true" model). All outcomes/hypotheses have the potential to happen but with different probabilities. This puts the Bayesian approach in stark contrast to the Classical approach where the goal of statistical analysis is to find the true model that presumably generates the observed data.

How can we get $P(\theta|y)$? As the Bayes Theorem shows above, we need three different probabilities, $P(\theta), P(y|\theta)$, and $P(y)$.

Prior distribution

$P(\theta)$ is called the prior distribution of θ and represents the prior knowledge about our hypotheses before we observe data. The issue is that it is often difficult to agree on universally accepted prior belief (or objective prior) for statistical problems at hand even though different prior specifications can lead to different statistical inference in the end. Many people find it rather uncomfortable with this "subjective" aspect of the prior belief. However, it has become clear that many classical statistical techniques as well implicitly use prior belief in one way or another as is in the case of "regularization." At a more fundamental level, every single step in the statistical modeling, regardless of whether classical or Bayesian, comes from the non-objective grounds such as subjective belief, widely-used conventions, or computational considerations. In a sense, the prior is part of the model, or the part of the statistical assumptions. It does not have to be label as "subjective" or "objective." Finally, it is worthwhile to mention that it has become clear that prior assignment of probability is extremely useful when dealing with highly complex models.

Likelihood and posterior

$P(y|\theta)$ is the likelihood function as Equation 6. It is often called a "model" since it connects the hypotheses and data by giving the probability of observed data according to hypotheses.[6] Note that the product of prior and likelihood is not a proper probability distribution. Therefore, we need a normalizing constant to make $P(\theta)P(y|\theta)$ sum up to one with respect to θ. The summation of $P(\theta)P(y|\theta)$ is just $P(y)$ by the law of total probability. That is, deviding $P(\theta)P(y|\theta)$ by $P(y)$ makes the posterior distribution, $P(\theta|y)$, a proper probability distribution.[7] With these three probabilities, we can construct the posterior probability distribution of our hypotheses. The role of these three probabilities will be more clear in the examples we will discuss throughout the course.

3.3 Example

To close on the topic of competing methods, let's repeat the same coin-tossing example as above using the Bayesian method. The Bayesian treatment of the Binomial trial is very similar to that of the MLE since they share the same likelihood function. However, in the Bayesian framework, i) there is a prior distribution for the parameter p and ii) the estimation result comes in the form of a probability distribution (the posterior distribution), not as a point estimate. As before, the likelihood of p is the binomial distribution with 5 trials and the parameter p to be estimated while the prior needs to be determined:

$$\begin{aligned} x &\sim \text{Binomial}(n = 5, p) \\ p &\sim \text{to be determined} \end{aligned}$$

[6] This convention that we call the likelihood a model is somewhat unfortunate since it gives a wrong impression that the prior is external to the model. As we discussed in the prior section, the prior distribution is a statistical assumption and thus is part of the modeling procedure.

[7] $P(y)$, which is often called "evidence" plays a very important role in model selection from the information-theoretic perspective.

Beta prior

Depending on what priors we use, the posterior distribution (as the normalized product of the likelihood and prior) will be determined. To highlight the impact of prior distributions, let us try different priors and how this makes difference in the posterior distributions of p. Note that the prior on p needs to be bounded between 0 and 1 since p is the "probability" (of landing heads). One of the most widely used prior for p is the Beta distribution.[8] The Beta distribution is a continuous probability distribution defined on the interval $[0, 1]$ with the following PDF with two positive shape parameters $\alpha > 0$ and $\beta > 0$:

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

where $\Gamma(.)$ is the Gamma function, $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\,dx$. The expected value and the variance are $\frac{\alpha}{\alpha+\beta}$ and $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$, respectively. Figure 9 shows the Beta distribution with varying parameter values of α and β. Note that Beta$(1, 1)$ is equivalent to the uniform distribution between 0 and 1, as shown in the black line in the left panel.

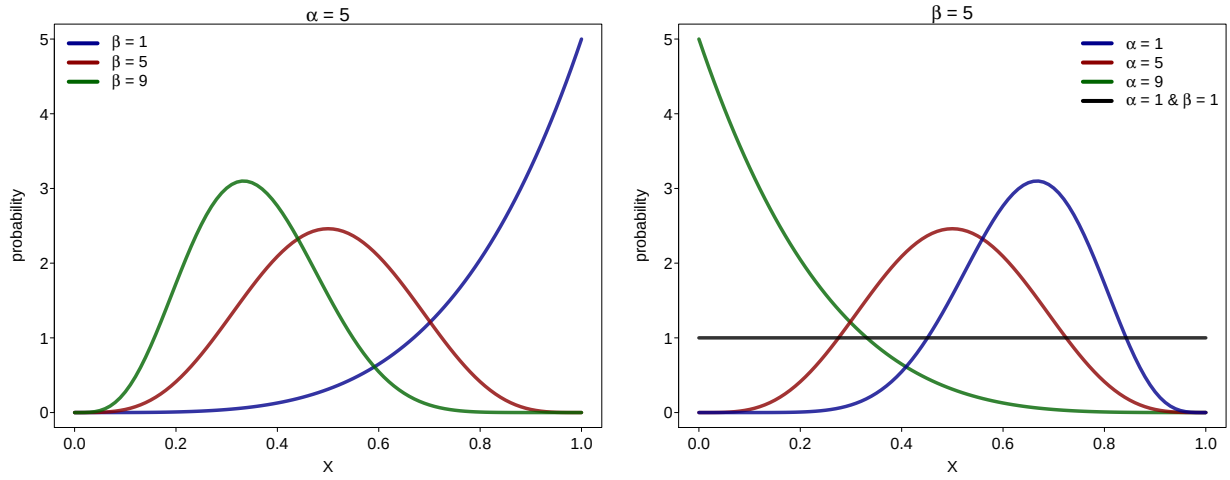


Figure 9: Beta probability density function with varying parameters.

Posterior as a compromise between likelihood and prior

Now, we can evaluate the impact of different priors on the posterior distribution. We choose four priors: i) Beta(1,1), which is a uniform distribution between 0 and 1, ii) Beta(5,5), a symmetric distribution centered at 0.5, iii) Beta(2,5), a right-skewed distribution whose mean is around 0.3, and vi) Beta(5,2), a left-skewed distribution whose mean is around 0.7. Figure 10 visualizes these priors (green) along with the resulting posterior distributions (red). The posterior distribution with the uniform prior is in black in the top-left panel. The likelihood function is added as a reference point in grey. The dotted vertical line represents the mode of each of the distributions (the peak of the distribution).

The impact of the prior is as predicted. The posterior with the uniform prior is exactly the same as the Binomial likelihood in Figure 8. This is generally true and the Bayesian estimation

[8] As we will discuss later in the course, the Beta distribution is the *conjugate* prior probability distribution for the binomial model meaning that the posterior distribution yields the same functional form of the prior.

with a uniform prior (or more broadly non-informative prior) is equivalent to the MLE. The symmetric prior, B(5,5), moves the posterior distribution slightly left to the likelihood function. The right-skewed prior, B(2,5), moves it to further left. Finally, the left-skewed prior, B(5,2), moves the posterior to the right. In all cases, the posterior distribution is a *compromise* between the prior distribution and the likelihood function.

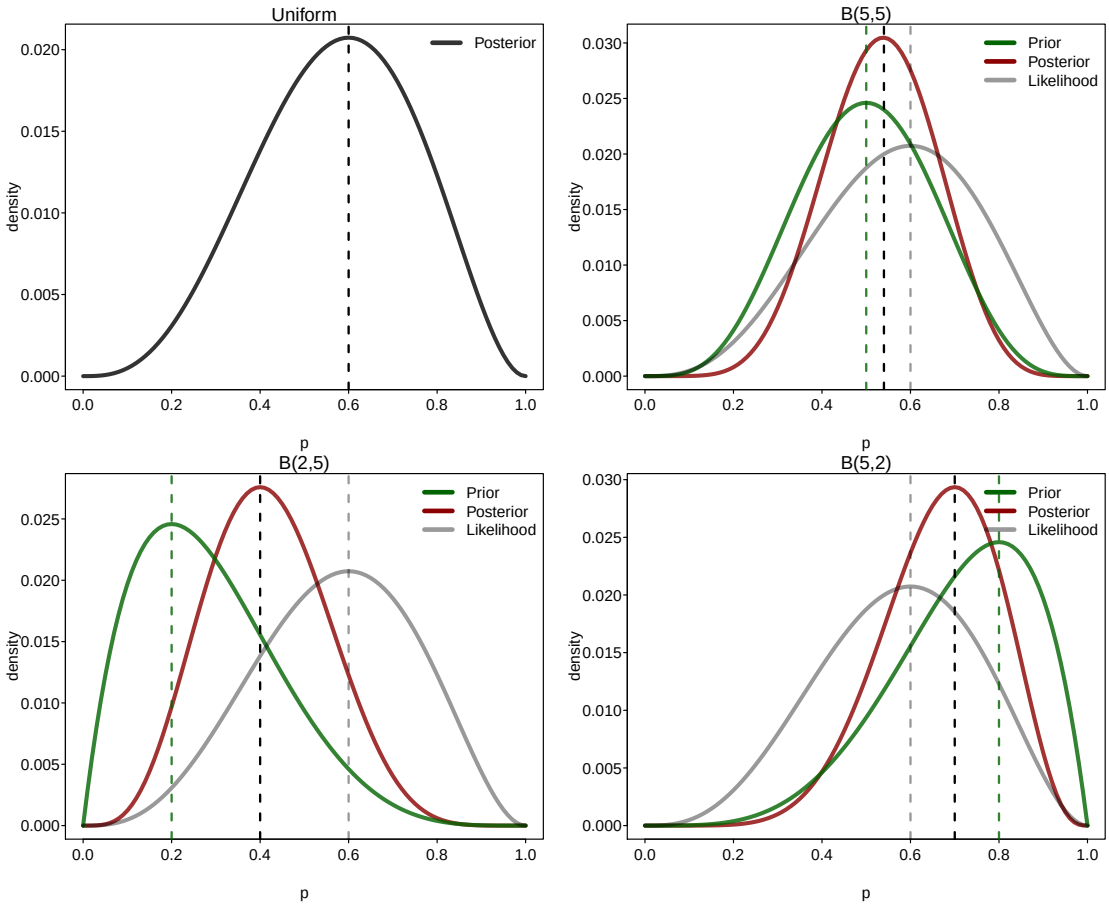


Figure 10: Posterior distribution of the probability of landing a head with different priors.

It is worthwhile to emphasize again that the estimation result comes in the form of a probability distribution not as a point estimate. This posterior distribution is different from the classical sampling distribution from which the confidence interval is constructed. In the sampling distribution, the given interval (confidence interval) is not the interval of the potential values of the quantity of our interest, p. Rather, the confidence interval is the range where the true value of the quantity of interest (p in our case) lies with a pre-defined confidence level when we repeat the same experiment many times. In contrast, the interval in the posterior distribution, which is often called the *uncertainty interval* or *credible interval* is indeed the interval of all possible values of the quantity of interest. The posterior distribution just assigns relative weights to each of these outcomes and, by doing so, provides information as to which hypothesis (among many possible outcomes of p, for example) is more consistent with our data and prior knowledge.

Beta posterior

The same Bayesian estimation procedure with a Beta prior can be illustrated analytically. With the prior on p, $p \sim \text{Beta}(\alpha, \beta)$, and ignoring the constant,

$$\begin{aligned} p(p|x) &\propto p(p|x)p(p) \\ &\propto p^x(1-p)^{n-x}p^{\alpha-1}(1-p)^{\beta-1} \\ &= p^{x+\alpha-1}(1-p)^{n-x+\beta-1} \end{aligned}$$

This shows that the posterior distribution of p is another beta distribution $p|x \sim \text{Beta}(\alpha + x, \beta + n - x)$, the same functional form of the prior $p \sim \text{Beta}(\alpha, \beta)$, with updated parameters.

However, many modern statistical problems are highly complex and it is extremely hard to find the set of priors leading to analytically solvable posterior distributions. Instead, we need some simulation methods to approximate the posterior distribution. Luckily, advances in computations and algorithms allow us to numerically obtain the posterior distributions like in Figure 10, so the statistical analysis on the estimated values can be accomplished. We will discuss how this simulation method works with examples in the latter part of the course.

Chapter 4: Regression Analysis: Simple Linear Regression

Jangho Yang

v1.0

Contents

1 Examples of linear predictions

Since we have established a basic framework for statistical inference, let us address some of the key statistical problems, starting with linear regression analysis that addresses a linear relation between two or more variables. Before we start, let's take a look at two examples of a linear relationship in data.

1.1 Income inequality and voting pattern

Figure 1 shows a linear relationship between the voter's income level and the voting tendency toward Republicans among different ethnic groups in the 2008 US presidential election (Gelman et al. 2010). There is a clear linear pattern among all different ethnic groups suggesting that the richer you are, the more likely you will vote for Republicans. Also, note a significant variation of the overall voting share of the Republican candidate across ethnic groups. For example, black people rarely vote for the Republican across all income groups. This group effect/variation in the linear relationship will be discussed in Topic 5 where we will introduce a multiple linear regression model.

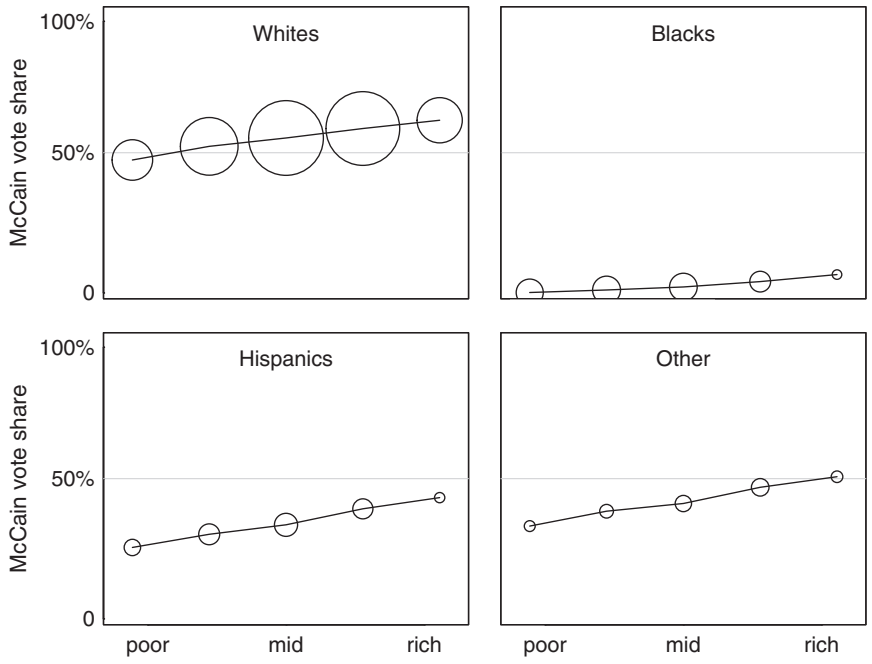


Figure 1: Voting pattern vs. income level from Gelman et al. (2010) Republican John McCain's vote share during the 2008 United States presidential election versus income level, among different ethnic groups (Pew Research Center Polls). The size of each circle is proportional to the number of voters in the category.

1.2 Corruption and economic growth

Figure 2 shows a relationship between corruption perceptions index (CPI) and the GDP per capita at a country level (Podobnik et al. 2008). There is a clear linear association between these two variables, suggesting that the more corrupt a country is, the slower its economic growth. It is worthwhile to mention that this linear association itself does not tell much about the specific

dynamics of how this relationship could happen. It is however a good starting point for such further analysis. For example, the authors of this paper took a further look at the relationship between foreign direct investment and the corruption level and found that less corrupt countries receive more foreign investments, which could lead to more economic growth.

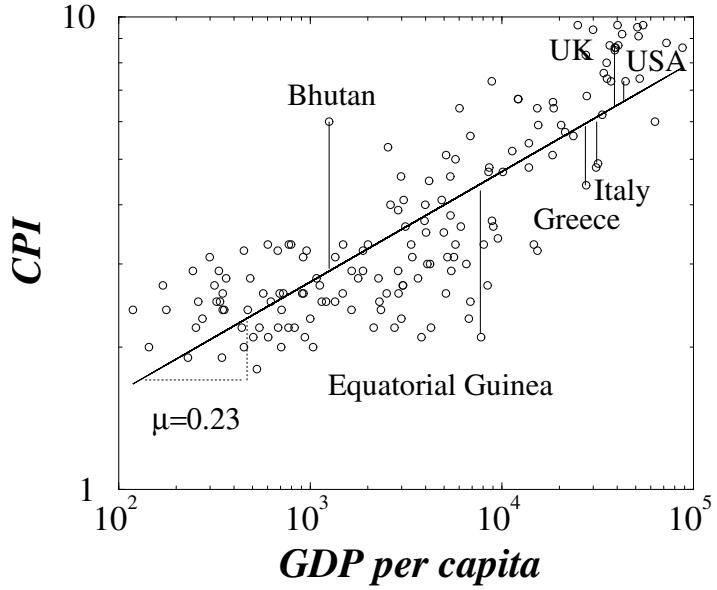


Figure 2: Corruption perceptions index (CPI) vs. economic growth from Podobnik et al. (2008). For the CPI, the lower the number, the more corrupt the country is.

2 Simple linear regression model: one predictor and one response variable

Regression analysis addresses particular statistical hypotheses about the relations/associations between variables. In this section, we will examine hypotheses about the *linear* relations only. This is obviously a tremendous simplification since there are infinitely many more non-linear relations than the linear ones in the real-world. However, linear regression models can serve as a very good starting point for a systemic understanding of the relationship between variables without sacrificing computational efficiency.

Assume that we have a sample of n observations on two variables, which we denote by x and y. The standard set up selects an *outcome variable* (or *dependent/response variable*) y and makes it a linear function of a *predictor* (*independent variable*) x:

$$y = \alpha + \beta x + \varepsilon, \tag{1}$$

where α and β are called *intercept* and *slope* parameters(or coefficients) respectively. This is because α is a y-intercept that shifts the value at $x = 0$ up and down on an $x - y$ plane, while β represents the slope of the linear line, which represents the *marginal increase* of y when x changes by one unit, or the marginal impact of x on y, dy/dx. Note that this geometric interpretation is not entirely straightforward when we have higher-order terms in

the linear equation or when we have multivariate relationships, but it gives good enough intuition for a simple linear relationship between two variables. Before we get to the *error term* ε, let's take a quick look at one example of a linear relation between x and y in the left panel of Figure 3. It is clear from the figure that there is a linear relation between x and y. Then, let's draw some random linear lines over the same data points in the right panel of Figure 3.

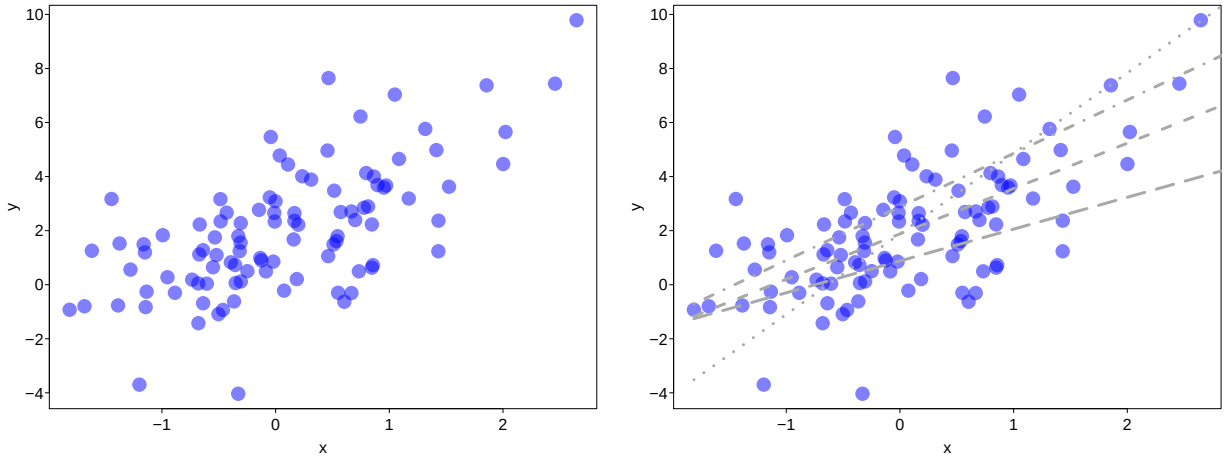


Figure 3: Linear relation between x an y. The right hand figure adds random linear lines to the scatterplot.

Which line best summarizes the linear relationship between the two variables? What are the criteria for your judgment? Here comes the error term $\varepsilon = y - \alpha - \beta x$, which plays a very important role in providing some criteria for the fit of the linear line. The standard set up for the error term is to assume that the errors (or disturbances) are independent and identically distributed (iid) and follow a normal distribution with mean zero:

$$\varepsilon \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2) \tag{2}$$

Setting aside the issue of how and why this normality assumption makes sense for now, let's first examine what this conventional specification of the error term implies. First, the zero-mean assumption means that given the x value, we want the data points to be distributed around the linear line symmetrically. That is, we do not want the linear line to be biased to either its left or right side. Second, we want the scale parameter σ to represent the degree of dispersion of data points away from the linear line. The larger σ is, the larger the dispersion. Figure 4 visualizes these two properties of the normality assumption.

Before we move to the estimation of this simple linear model, let us briefly discuss some justifications/falsifications for the normality assumption. The normal distribution is summarized by the first and the second moments of the random variable (mean and the variance). Therefore, if researchers want to avoid the computational burdens of taking into account higher sample moments in their analysis and want to approximate the data by using the first two moments, the normal distribution is the natural choice.[1] Second, the normal distribution is a meta distribution of the aggregate sum (or the average) of the random variables. That is, when we draw n random

[1]It can be mathematically shown that the normal distribution is the *most likely distribution* or the *maximum entropy distribution* when the random variable is constrained by the first two moments in the real line.

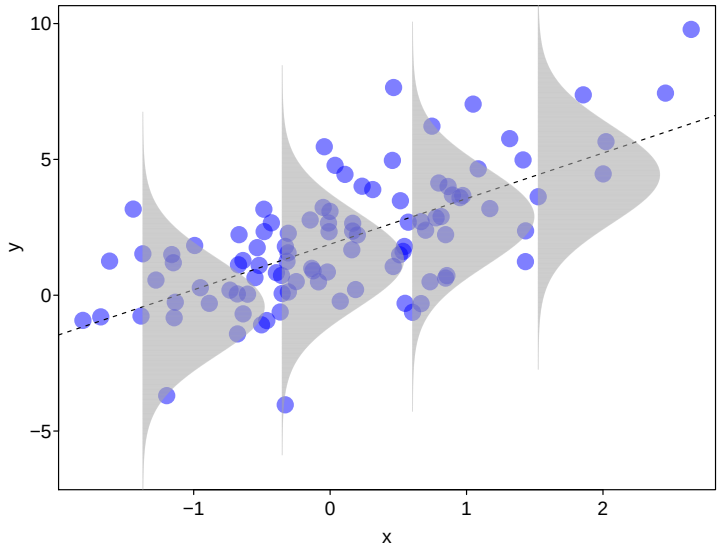


Figure 4: Linear relation between x an y with normal error term.

numbers from any distributions and take the average of them, and we repeat this many times, these averages will converge to a normal distribution. The more numbers we draw ($n \uparrow$), the more it looks like a Normal. Therefore, if the researchers are interested in the variable that presumably comes from the aggregation of underlying components, such as a firm's productivity as an average of productivities of its employees, the Normal assumption can be justified.

The key drawback of a normal distribution is that it has a *thin tail*, meaning that it cannot capture observations that are far away from the mean. To get some intuition, suppose we have some observations of stock returns (fake data) clustered around the mean so that we want to use a normal distribution to describe it, which is shown in Figure 5. We have two normal distributions. The first normal distribution in black is fitted only using values between -15% and 15% while the second one in red uses all the data points.

It is clear from the figure that the first normal explains the data around the mean quite well but has practically zero probability for stock market return below -15% and above 15%. The second normal seems to encompass the extreme values with a higher scale parameter but poorly predicts the data points around the mean. This exercise suggests that the normal distribution is not a good model when the data includes many extreme values. In many real life data, we observe considerable extreme values (black swans). For example, stock prices tend to be extremely volatile historically so that its time series is punctuated by a series of dramatic booms and busts. That is, -15% or 15% market returns (at least for individual stocks) are quite common. When we have extreme values in the data, researchers use *heavy-tail* distributions to account for extreme values far away from the mean, such as Student-t distribution and Cauchy distribution.[2]

[2]See Taleb (2007) for a comprehensive discussion on extreme values and fat-tail distributions

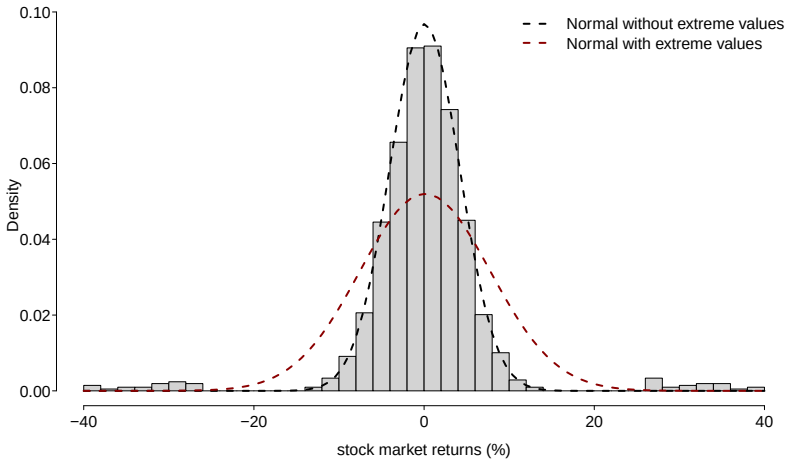


Figure 5: Distribution of stock returns with two different normal fits with and without extreme values.

3 Estimation and inference methods

Now that we have discussed the basic concepts and model specification of a linear regression model, let us turn our attention to how to estimate this model from various methodological points of view.

3.1 Ordinary Least Squares and classical inference

OLS estimation

One of the most standard estimation methods for linear regression is the Ordinary Least Squares (OLS) estimation. Loosely speaking, it is based on a very simple and intuitive idea that we want our prediction of data points as close to actual data points as possible. In other words, the OLS estimation gives the estimator of a linear line that minimizes the distance between predicted values and actual observations. Let's unpack what this means by working out some equations.

Formally, the OLS estimation minimizes *the sum of the squares* of the differences between the observed dependent variable and its predicted values, often called *residuals*. Let $\hat{\alpha}$ and $\hat{\beta}$ denote the estimator of α and β. Then, the predicted values of the dependent variable on the linear line, which we denote by $\hat{y}$, is

$$\hat{y} = \hat{\alpha} + \hat{\beta}x \tag{3}$$

What we want is to find $\hat{\alpha}$ and $\hat{\beta}$ such that

$$\sum_{i}^{n}(y_i - \hat{y}_i)^2 = \sum_{i}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 \tag{4}$$

is the minimum. To find $\hat{\alpha}$ and $\hat{\beta}$, we need to solve the following optimization problem:[3]

$$\arg\min_{\alpha,\beta} \sum_{i}^{n}(y_i - \alpha - \beta x_i)^2 \tag{5}$$

[3]Note that $\arg\min f(x)$ represents the value of x at which the maximum of $f(x)$ is attained.

The solution to this optimization problem is

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{6}$$

$$\hat{\beta} = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})^2} = \frac{\text{Cov}(x, y)}{\text{Var}(x)} \tag{7}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of x and y. The proof of this result will be shown in Topic 5 in a more general framework. This result shows that the OLS estimators for α and β are functions of the sample means, variances, and covariances of x and y.[4]

Classical inference

From here, the goal of the classical statistics is to find the correct sampling distribution of these estimators to construct a confidence interval (See Topic 3). Using some elementary algebra, it can be shown that the sampling distributions for $\hat{\alpha}$ and $\hat{\beta}$ are the following:

$$\hat{\alpha} \sim \text{N}\left(\alpha, \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{x}^2}{\sum_i^n (x_i - \bar{x})^2}\right) \tag{8}$$

$$\hat{\beta} \sim \text{N}\left(\beta, \frac{\sigma^2}{\sum_i^n (x_i - \bar{x})^2}\right) \tag{9}$$

This result shows that the sampling distribution of the OLS estimator is a normal distribution. However, this only holds when we know σ and do not have to estimate it along with $\hat{\alpha}$ and $\hat{\beta}$. When σ is unknown, we need to estimate it as well, which results in a different form of the sampling distributions for $\hat{\alpha}$ and $\hat{\beta}$. The unbiased estimator for σ for a simple linear regression with one intercept and one slope is

$$\hat{\sigma} = \sqrt{\frac{1}{n-2}\sum_i^n (y_i - \hat{y})^2} \tag{10}$$

In other words, we use the sum of the squares of the residuals corrected by two degrees of freedom for our estimator for σ. The correction of $n-2$ is necessary in order to correct for a downward estimation of σ when using residuals.[5]

With unknown σ that needs to be estimated using a finite sample, there is more uncertainty in the sample distributions of $\hat{\alpha}$ and $\hat{\beta}$. As we discussed in Topic 3, it turns out that when the sample size is not large enough, a Student-t distribution is the correct sampling distribution for $\hat{\alpha}$ and $\hat{\beta}$ as follows

$$\hat{\alpha} \sim \text{t}_{n-2}\left(\alpha, \frac{\hat{\sigma}^2}{n} + \frac{\hat{\sigma}^2 \bar{x}^2}{\sum_i^n (x_i - \bar{x})^2}\right) \tag{11}$$

$$\hat{\beta} \sim \text{t}_{n-2}\left(\beta, \frac{\hat{\sigma}^2}{\sum_i^n (x_i - \bar{x})^2}\right) \tag{12}$$

[4]These estimators are often called *the best linear unbiased estimator* (BLUE), meaning that they have the lowest sampling variance among unbiased estimators, which can be shown by the *Gauss–Markov theorem*.

[5]This is a subtle point that requires some elaboration. When estimating the standard deviation of the normally distributed data with an unknown mean, we use a sample standard deviation corrected by $n-1$, which is often called *Bassel's correction*. The intuition behind this correction is that the sample standard deviation, which represents the deviation of observations from their mean (the sample mean), underestimates the true standard deviation since it doesn't account for the potential deviation of the sample mean from the true mean. To correct for this source of bias from one unknown variable that leads to a downward estimation of the standard deviation, we need $n-1$ correction when using the sample standard deviation. The same logic goes for the linear regression exercise. Here, the mean (of the normal distribution) consists of two variables, the intercept and the slope coefficients, meaning that we have two sources of bias. To correct for this, we need $n-2$ correction, when using the sample residuals for the estimation of σ.

Using these sampling distributions, one can easily construct the confidence interval for α and β.

Example

Let's now look at an example to see how to implement the classical method of simple linear regression. The data we use are from Figure 3. We have 100 observations of x and y variables as shown in Table 1.

Table 1: Example Data of x and y

Obs.	x	y
1	-1.81	-0.93
2	-1.68	-0.80
3	-1.62	1.26
⋮	⋮	⋮
99	2.46	7.44
100	2.65	9.79

Using Eqs 7 and 10, we can calculate the estimators for α, β, and σ as follows.

$$\begin{aligned} \hat{\alpha} &= 1.88 \\ \hat{\beta} &= 1.68 \\ \hat{\sigma} &= 1.79 \end{aligned}$$

Since we have 100 observations, it is safe to use a normal distribution as a sampling distribution to construct the confidence interval. Using Equation 13, we have

$$\begin{aligned} \hat{\alpha} &\sim \mathrm{N}\left(\alpha, 0.13^2\right) \\ \hat{\beta} &\sim \mathrm{N}\left(\beta, 0.15^2\right) \end{aligned}$$

Using the sampling distributions, we can construct confidence intervals. Figure 6 visualizes the regression result with a fitted regression line and confidence intervals of different α-levels.

3.2 Maximum likelihood estimation

The MLE is a statistical procedure that finds the point estimate of parameters by maximizing a likelihood function associated with them (See Topic 3). Then, what is the likelihood function for the simple linear regression? Assuming that the errors follow a normal distribution with zero mean, we can rewrite the simple linear regression as follows

$$y \sim \mathrm{N}(\alpha + \beta x, \sigma^2) \tag{13}$$

This is another way of saying that the dependent variable y is normally distributed with standard deviation σ along the regression line defined by $\alpha + \beta x$ on the $x - y$ plane. Remember that the likelihood function is the probability of data given the hypotheses $P(y|\theta)$. Therefore, the likelihood function for the simple linear regression, $\widehat{L}_n(\alpha, \beta, \sigma\,|y)$ can be written as follows

$$\begin{aligned} \widehat{L}_n(\alpha, \beta, \sigma\,|y) &= \prod_i^n P(y_i|x_i, \alpha, \beta, \sigma) \\ &= \prod_i^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2}\right) \end{aligned} \tag{14}$$

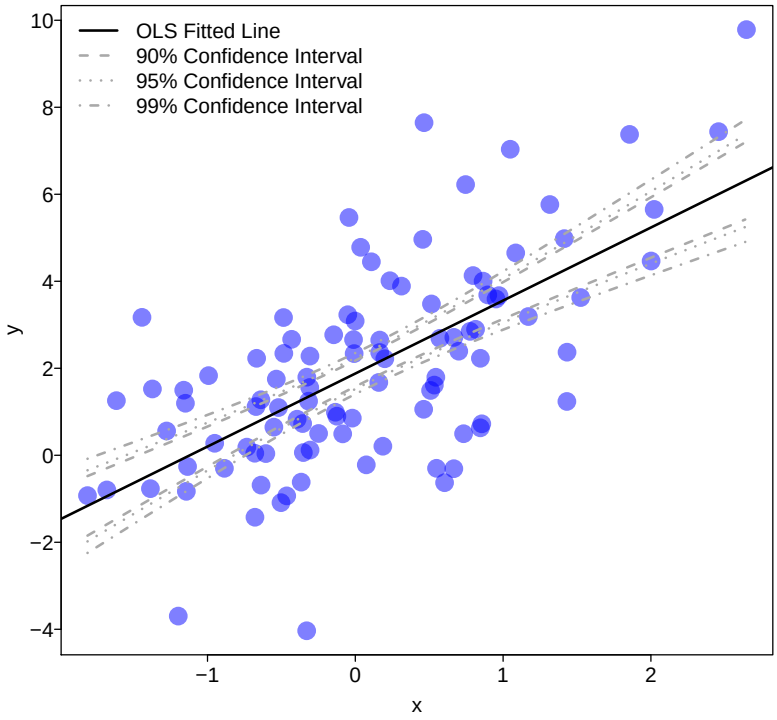


Figure 6: OLS regression line with confidence intervals.

The log-likelihood function for the linear regression is written as follows

$$\widehat{l}_n(\alpha, \beta, \sigma \,|y) = \log \prod_i^n P(y_i|x_i, \alpha, \beta, \sigma) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_i^n (y_i - \alpha - \beta x_i)^2 \tag{15}$$

Therefore, the MLE for the simple linear regression boils down to solving the following optimization problem

$$\hat{\alpha}, \hat{\beta}, \hat{\sigma} = \arg\max_{\alpha,\beta,\sigma} \widehat{l}_n(\alpha, \beta, \sigma \,|y) \tag{16}$$

where $\widehat{l}_n$ represents a log-likelihood function. The solution for each of the parameters is

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{17}$$

$$\hat{\beta} = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})^2} = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)} \tag{18}$$

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_i^n (y_i - \hat{y})^2} \tag{19}$$

The result shows that the MLE estimators for $\hat{\alpha}$ and $\hat{\beta}$ are exactly the same as the OLS estimators. In contrast, the MLE estimator for σ doesn't include $n-2$ correction and is biased. However, this estimator is consistent, meaning that as the sample size increases, it converges to the true σ.

How about the sampling distribution for the MLE estimators? As expected, the sampling distributions for $\hat{\alpha}$ and $\hat{\beta}$ are exactly the same as the OLS sampling distributions as follows

$$\hat{\alpha} \sim \mathrm{N}\left(\alpha, \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{x}^2}{\sum_i^n (x_i - \bar{x})^2}\right) \tag{20}$$

$$\hat{\beta} \sim \mathrm{N}\left(\beta, \frac{\sigma^2}{\sum_i^n (x_i - \bar{x})^2}\right) \tag{21}$$

Even though we are not showing here formally, it is worthwhile to note that the MLE method allows us to construct the sampling distribution of σ^2 as well.

We are skipping the example of the MLE for a simple regression since the results will be the same as those from the above. MLE implementation will be discussed in great detail in the next topics.

3.3 Bayesian methods

The Bayesian treatment of the simple regression analysis is very similar to that of the MLE since they both share the same likelihood function. As we discussed in Topic 3, in the Bayesian framework, i) there is a prior distribution for the regression parameters, $P(\theta) = P(\alpha, \beta, \sigma)$, and ii) the estimation result comes in the form of a probability distribution (the posterior distribution).

There are multiple ways of approaching the simple linear regression analysis from the Bayesian perspective depending on what types of priors we want to use. In fact, when we use uninformative priors, the Bayesian method is equivalent to the MLE, and therefore the MLE estimators (and consequently, OLS estimators) are the same as the posterior mean of α and β.[6] As we discussed in Topic 3, Bayesian statisticians often use a *conjugate prior*, a particular type of prior distribution that makes the posterior distribution yield the same functional form as the prior. It can be shown that the conjugate prior on the coefficients in the linear regression with the normally distributed errors is a normal distribution. That is, if we use a Normal prior on α and β, the posterior distribution of α and β itself will be a normal distribution. The use of a conjugate prior is mathematically elegant but only works for relatively simple models such as the simple linear regression in our exercise. For more complex problems, it is often not easy to find conjugate priors due to mathematical difficulty. Therefore, to highlight the flexibility of the Bayesian method, we will not limit our discussion to the conjugate priors and show that Bayesian statistical inference can be done without worrying about the mathematical difficulty of solving the posterior distributions.

Example

To give an intuition of how to set up a Bayesian model for a simple linear regression, let's repeat the example we used above using the Bayesian method. Remember that the goal of the Bayesian estimation is to get the posterior distribution, which is essentially the probability of all possible values of the coefficient given data, $P(\theta|y)$. According to Bayes' theorem, we need three probabilities to get the posterior, namely, $P(\theta), P(y|\theta)$, and $P(y)$. Ignoring the normalizing constant, $P(y)$, $P(\theta|y)$ is proportional to the product of $P(\theta)P(y|\theta)$. Therefore, in our linear

[6] With a bit of math, it can be shown that the posterior distribution of σ^2 follows an inverse χ^2 distribution with $n-2$ degree of freedom.

example exercise, we can write the posterior as follows

$$P(\alpha, \beta, \sigma | y, x) \quad \propto P(\alpha, \beta, \sigma) P(y | x, \alpha, \beta, \sigma) \tag{22}$$

where $P(\alpha, \beta, \sigma)$ is a joint prior distribution on α, β, and σ while $P(y|x, \alpha, \beta, \sigma)$ is a likelihood function.[7][8] As we discussed above, the likelihood function for a simple linear regression can be expressed as $y \sim \mathrm{N}(\alpha + \beta x, \sigma^2)$. Then, Bayesian priors distributions are applied to three unknowns, α, β, and σ.

$$\begin{aligned} y &\sim \text{Normal}(\alpha + \beta x, \sigma^2) && (23)\\ \alpha &\sim \text{to be determined} && (24)\\ \beta &\sim \text{to be determined} && (25)\\ \sigma &\sim \text{to be determined} && (26) \end{aligned}$$

Depending on what priors we use, the posterior distribution will be determined. Let's start with non-informative priors on all three coefficients. Suppose α and β can take all values in the real line, then a prior distribution shouldn't be bounded. This is not the case for σ because the standard deviation is always positive. For this reason, a prior distribution on σ needs to have only positive support. For this exercise, we will use a normal distribution as a prior on α and β and an exponential distribution as a prior σ. To make this prior distribution uninformative, we need to assign an almost equal probability to each of all possible values of the coefficients and therefore, make the distribution as dispersed as possible.[9] For a normal distribution, this can be done by setting the scale parameter to be high. For an Exponential distribution, the rate parameter needs to be small. (See Topic 1 for discussion on the normal and exponential distribution.)

Let's set the standard deviation of a normal distribution to be 10 and the rate of the Exponential to be 0.1 as follows.

$$\begin{aligned} y &\sim (\alpha + \beta x, \sigma^2) && (27)\\ \alpha &\sim \mathrm{N}(0,\ 10) && (28)\\ \beta &\sim \mathrm{N}(0,\ 10) && (29)\\ \sigma &\sim \text{Exponential}(0.1) && (30) \end{aligned}$$

Then, our prior distributions are visualized in Figure 7.

Since the prior distribution covers a wide range of values, it correctly reflects our prior ignorance. We can check this by simulating data for y given these priors, which is called *prior predictive check*. Figure 8 shows prior predictive lines over the original data generated from the

[7] Assuming independence of prior distributions, $P(\alpha, \beta, \sigma) = P(\alpha)P(\beta)P(\sigma)$.

[8] Note that our simple regression analysis assumes the independent variable x to be known (or non-random). A full probabilistic model for the linear regression involves a joint likelihood $P(x, y|\theta)$. This means that the distribution of x, which we assume to depend on some arbitrary set of parameters *phi*, can feed into the likelihood function of the linear regression. However, under a weak assumption that $P(x|\phi)$ and $P(y|x, \theta)$ are independent, it can be shown that we can ignore the distribution of x and can work directly with $P(y|x, \theta)$.

[9] If we want to assign the exactly equal probability to all possible values, we can use a uniform distribution as a prior. However, when the potential values of coefficients are not bounded, the uniform distribution is an *improper prior* in the sense that the distribution does not integrate to one. When the prior is improper, it is sometimes hard to find a valid posterior distribution.

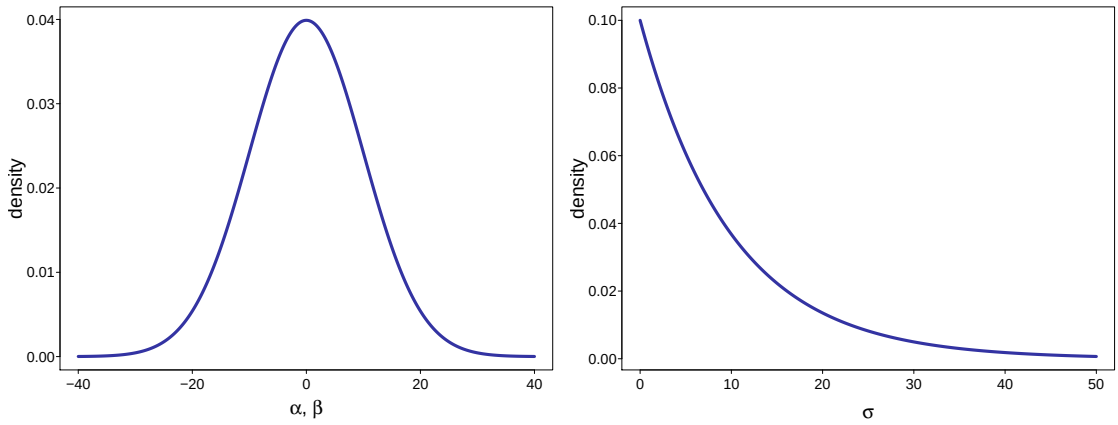


Figure 7: Normal and exponential priors with a wide support.

normal prior distributions of α and β. It clearly shows that the prior predictive lines are scattered all over the space with substantially larger ranges than the original data, implying that our prior distributions are not informative.

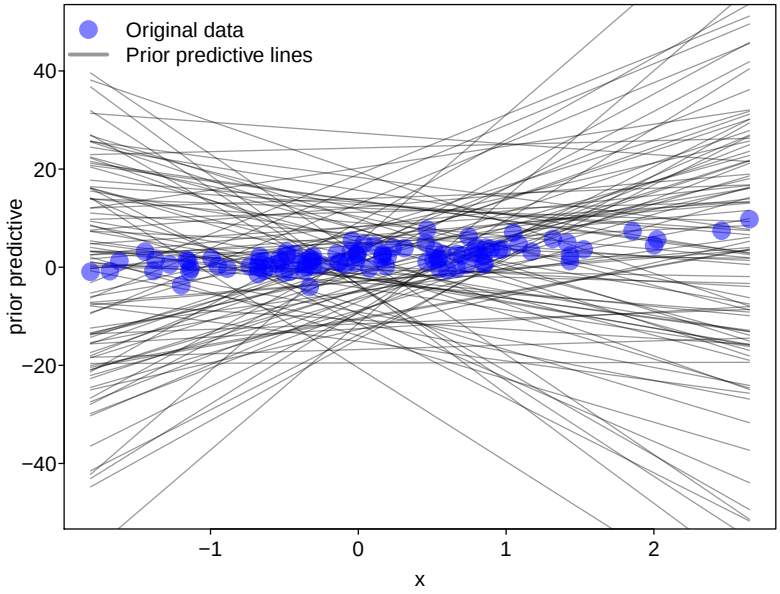


Figure 8: Prior predictive lines generated from the normal prior distribution of α and β.

As expected, the linear lines cover almost all spaces in the $x - y$ plane, meaning that our prior distributions are indeed uninformative.

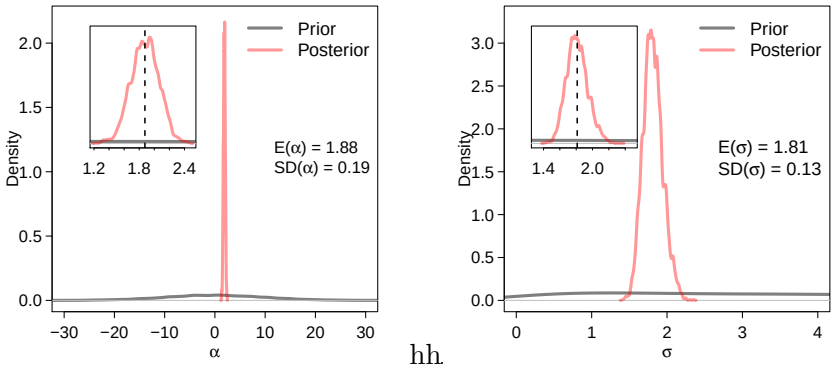
Now, how can we get the posterior distribution given these priors and the likelihood function? We can work out the math and derive a functional form of the posterior distribution. Again, this is only possible for a simple problem. For more complex problems, it might be extremely hard to derive a posterior distribution analytically. Therefore, Bayesian data analysis always

goes along with simulation methods.[10] The most well-established simulation algorithm that is used to approximate the posterior distribution is Markov Chain Monte Carlo (MCMC). With this simulation method, the posterior distribution for our simple linear regression model can be approximated. Table 2 shows the mean, standard deviation, and the 5%, 50% (median), and 95% quantile of the posterior distributions along with the convergence diagnostics, $\hat{R}$.[11]

Par	Mean	SD	5%	50%	95%	$\hat{R}$
α	1.88	0.19	1.57	1.88	2.17	1.00
β	1.67	0.20	1.34	1.68	2.01	1.00
σ	1.81	0.13	1.61	1.81	2.04	1.00

Table 2: Summary statistics of the posterior distributions.

Figure 9 compares the posterior and prior distributions of α, β, and σ.



hh

Figure 9: Posterior distributions of α, β, and σ with corresponding prior distributions.

Note that the estimation result for all coefficients, α, β, and σ, comes in the form of a probability distribution, making it easy and intuitive to express our uncertainty of the estimation.

4 Post-estimation evaluation

4.1 Residual analysis

Now that we have discussed the estimation of a simple linear regression model, we need to validate how good our estimation result is. In the examples above, we visually checked the regression fit just by looking at whether the regression line passes through the data points. This is a good starting point, but there are some other aspects of estimation results we need to examine.

One of the key assumptions behind our simple linear regression is that errors are iid and are normally distributed with mean zero and a constant standard deviation. We need to verify if this assumption is valid. To do this, we can use residuals, which can be considered estimates of the errors, and see if they follow the normality assumption. Since we know both y and $\hat{y}$, we can

[10] We can use a grid approximation or Laplace approximation to find the posterior but these methods are not tractable with a model with a large number of parameters.

[11] When $\hat{R}$ is smaller than 1.1, it is safe to say that the MCMC chains are properly mixed

calculate the residuals by calculating their differences $e = y - \hat{y}$. With this, we can check whether residuals have i) mean zero, ii) constant variance and iii) no correlation (from the iid assumption).

Figure 10 visualizes three regression results with different residual patterns. The first column shows the fitted regression line $\hat{y}$ over the data x with the residual line (a vertical line between the observation and the fitted line). The second column shows the distribution of the residuals, while the third column shows the regression residuals versus the fitted line, which is often called a *residual plot*.

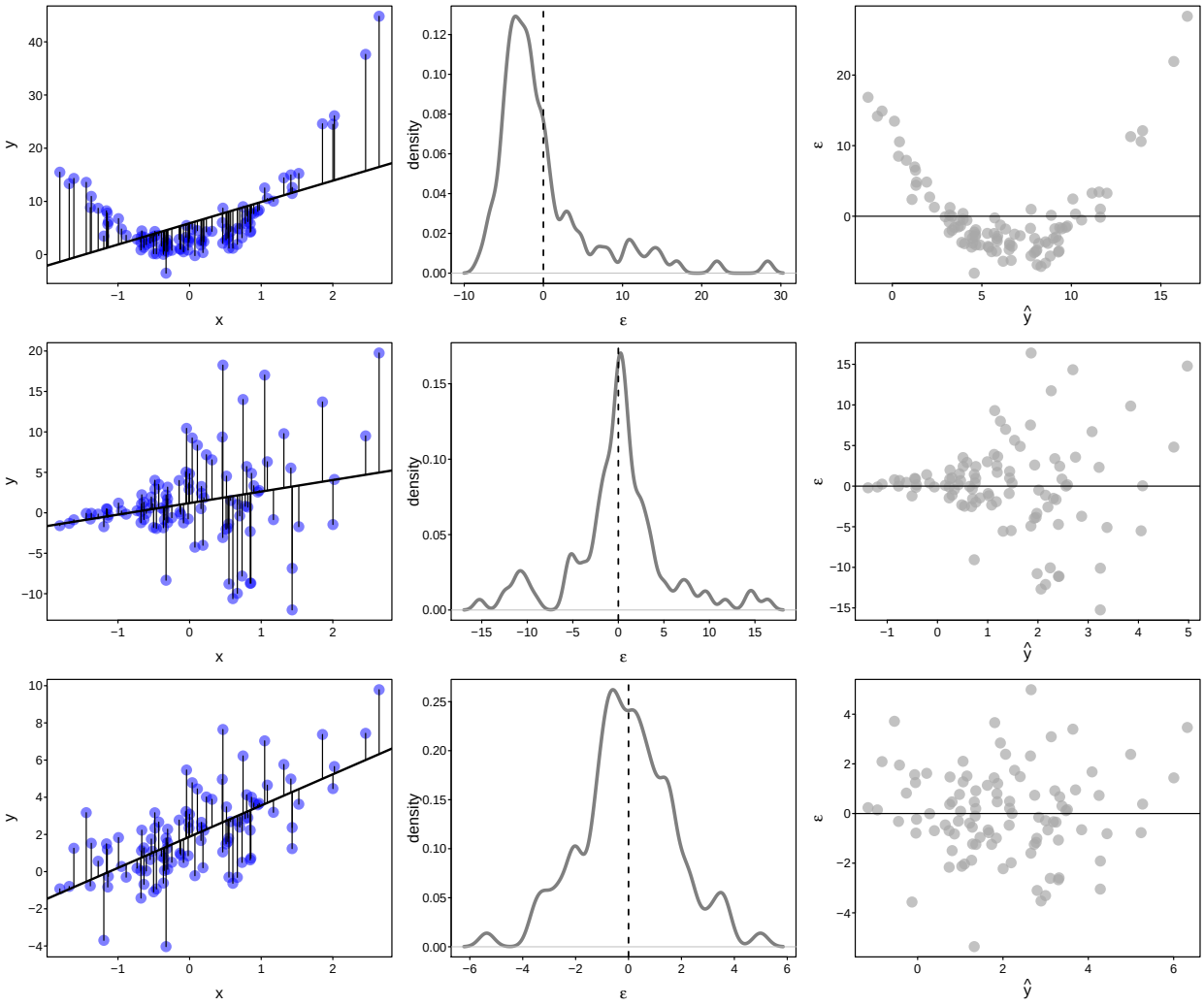


Figure 10: Three regression examples

The top row is a bad example of a simple linear regression since the observed data points are not centered around the estimated regression line. The regression line systematically overestimates y when x is very small or very large while it underestimates y when x is in the mid-range. This results in a correlation in the neighboring residuals, violating the independence assumption, as can be shown in the residual plot.

The second row shows another bad example. Even though the data points do seem to be centered around the regression line, there is inconsistency in how much the data points deviate from the regression line. As x increases, there is more deviation, implying that the assumption of a constant standard deviation (the assumption of identical distribution) is violated . This can be clearly seen in the residual plot that shows the increasing magnitude of residuals as x increases.

The last one shows a good example. The regression line passes through all the data points evenly so that the observed data points are centered around the estimated regression line with no changes in the degree of deviation. The residual plot exhibits consistently and evenly distributed residuals.

4.2 Goodness of fit

The residual plot shows whether the estimation result is consistent with the normality assumption of errors. However, it does not directly tell us much about the fit of the model itself, e.g. how well the model explains the observed data. There is a wide range of techniques that attempt to quantify how the observed data are different from the predicted values. These techniques are broadly called *goodness of fit test* or more broadly *model validation*. In this section, we will discuss a standard measure for goodness of fit in linear regression, namely, coefficient of determination.

Coefficient of determination

The coefficient of determination, better known as R^2, is defined as the proportion of the variance of data explained by the model. In other words, R^2 compares the degree of variability of y (put it differently, the prediction error of y) with and without the regression model in a particular manner.

There are different measures of variability, e.g. the sum of squares, absolute deviation, interquantile range, and R^2 uses the *sum of squares* as a measure of variability. The degree of variation in y before we use the model is just the sum of squares of y compared to its mean $\bar{y}$, which is the (unnormalized) variance of y. We call this measure the *total sum of squares*

$$SS_{\text{tot}} = \sum_i^n (y_i - \bar{y})^2 \tag{31}$$

In contrast, the degree of variation in y with the model is the sum of squares of y compared to its predicted value $\hat{y}$, which is the (unnormalized) variance of residuals. We call this the *residual sum of squares*.

$$SS_{\text{res}} = \sum_i^n (y_i - \hat{y}_i)^2 \tag{32}$$

Then, R^2 is defined as

$$R^2 = \frac{SS_{\text{tot}} - SS_{\text{res}}}{SS_{\text{tot}}} = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} \tag{33}$$

This is the proportion of the variance of y explained by the model. Alternatively, this is the fraction of the variation in y explained by x.[12] Two extreme cases are $R^2 = 0$ and $R^2 = 1$. When $R^2 = 0$,

[12]We can show that $\sum_i^n (y_i - \bar{y}) - \sum_i^n (y_i - \hat{y}_i) = \sum_i^n (\hat{y}_i - \bar{y})$, which we call *explained sum of squares* and denote SS_{reg}. SS_{reg} is the sum of square of the predicted value of y compared to the data mean, and can be interpreted as the (unnormalized) variance of the predicted value of y since OLS yields $\hat{y} = \mathrm{E}(y)$

this means that zero percent of the variance in y is explained by the explanatory variables. When $R^2 = 1$, 100% percent of the variance in y is explained by x, meaning the perfect fit. Normally, R^2 is calculated between 0 and 1. Figure 11 visualizes how to calculate R^2 with 4 data points.

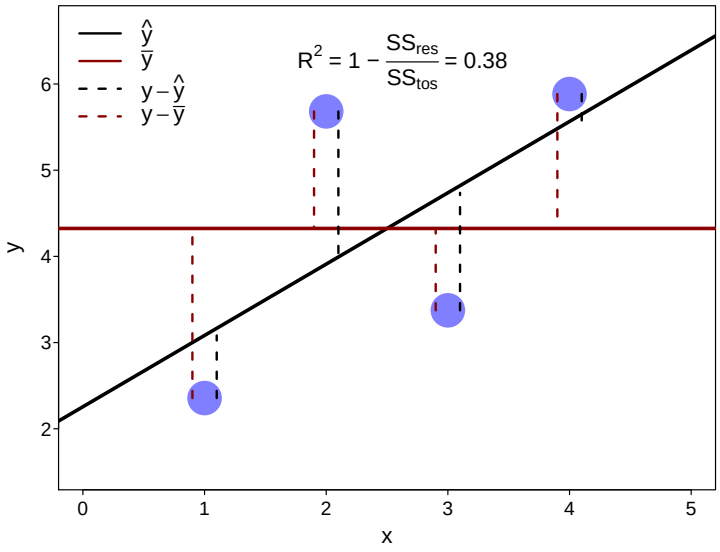


Figure 11: Visualization of R^2.

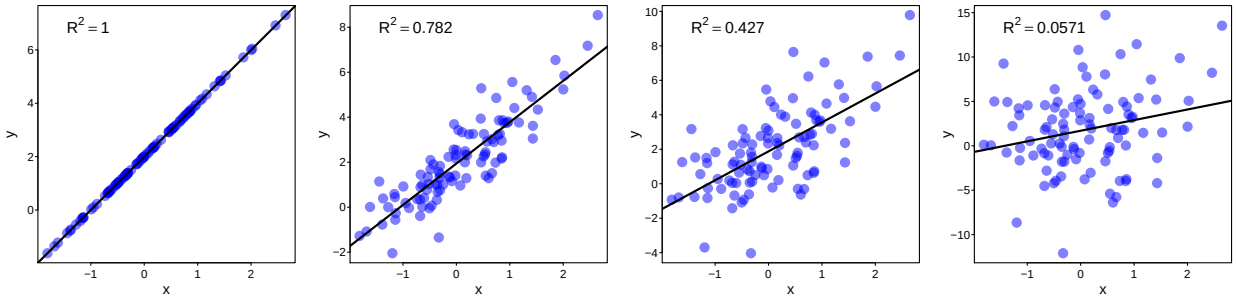


Figure 12: Regression results with R^2.

Figure 12 shows 4 different regression results with the corresponding R^2. The linear regression in the first panel has the best fit with $R^2 = 1$. R^2 gets smaller as the data become noisier. It is worthwhile to note that even though R^2 gives some information about the model fit, it cannot be used as a model comparison tool when comparing regression models with a different number of independent variables. This is because R^2 is an increasing function of the number of parameters, and therefore, adding more independent variables automatically increases R^2. We will come back to this point in Topic 6 where we discuss a multivariate regression model and *adjusted* R^2.[13]

[13]There is a Bayesian version of R^2. An advantage of using a Bayesian R^2 is that we can construct the probability distribution of R^2, providing useful information about the uncertainty of the model fit. See Gelman et al. (2019) for a detailed discussion on how to construct R^2 from the Bayesian perspective.

References

Gelman, A., Goodrich, B., Gabry, J. & Vehtari, A. (2019), 'R-squared for bayesian regression models', *The American Statistician* **73**(3), 307–309.

Gelman, A., Kenworthy, L. & Su, Y.-S. (2010), 'Income inequality and partisan voting in the united states', *Social Science Quarterly* pp. 1203–1219.

Podobnik, B., Shao, J., Njavro, D., Ivanov, P. C. & Stanley, H. E. (2008), 'Influence of corruption on economic growth rate and foreign investment', *The European Physical Journal B* **63**(4), 547–550.

Taleb, N. N. (2007), *The black swan: The impact of the highly improbable*, Vol. 2, Random house.

Chapter 5: Multiple Linear Regression Analysis

Jangho Yang

v1.0

Contents

1 Issues with a simple linear model

In Topic 4, we discussed a simple linear regression, a linear relation between one outcome variable and one predictor. We now turn our attention to a slightly more complex model, namely a *multiple linear regression* model. It is "multiple" because we use more than one predictor.

One might wonder why we have to use a multiple linear model instead of running multiple simple linear models since both will tell us how multiple predictors are correlated with an outcome variable. This intuition turns out to be misplaced. This section deals with some justifications of multiple linear regression models.

1.1 Simpson's paradox

Simpson's paradox is a statistical phenomenon where including another predictor reverses or nullifies the association between existing predictors and the outcome variable. That is, the coefficients of some predictors can be reversed or become zero when adding additional predictors. Let's take a couple of examples and discuss its implication with regard to the justification for a multiple regression model.

Amateur vs. professional climber

Suppose we have two rock climbers, A and B and there are two practice routes graded V4 and V8. V4 is not easy but one can do it with a bit of exercise, whereas V8 is quite hard and even professionals climbers sometimes fail. Both climbers have tried these two practice routes multiple times and recorded the success rates. They first recorded the overall success rate (# of success over # of trials) as is shown in the following table.

	Climber A	Climber B
Success rate	32% (19/60)	39% (120/305)

Climber B appears to have a higher overall success rate. Does this mean that Climber B performs better than Climber A? According to the overall success rate, the answer is yes. Now, let's look at the success rate by each route.

	Climber A	Climber B
V4 success	90% (9/10)	40% (120/300)
V8 success	20% (10/50)	0% (0/5)
Overall success	32% (19/60)	39% (120/305)

When looking at each of the practice routes, the success rates are reversed and Climber A outperforms B both in V4 and V8. That is, "conditionally on" each grade, Climber A has always higher success rates. So, what's happening? Climber A did have a lower overall success rate, but this happened because Climber A tried more V8 than V4 while Climber B tried substantially more V4 than V8.

College admission

The same logic can be applied to a real-life example of alleged discrimination in graduate school admission at UC Berkeley. The admission data for 1973 at UC Berkeley shows that male applicants had a higher chance to get admitted as shown in the table below

	Males	Females
Admission rate	44% (3738/8442)	35% (1494/4321)

According to this overall admission rate, one might conclude that relatively more males were admitted than females and there might have been some discrimination against female applicants. But again, if we look at individual departments, female candidates had a higher admission rate in a greater number of departments. Similar to the professional climber in the example above, females applications did have a lower admission rate, but this happened because females applied to the harder departments to get in.

Department	Men	Women
A	62% (512/825)	82% (89/108)
B	63% (313/560)	68% (17/25)
C	37% (120/325)	34% (202/593)
D	33% (138/417)	35% (131/375)
E	28% (53/191)	24% (94/393)
F	6% (22/373)	7% (24/341)
Overall rate	44% (3738/8442)	35% (1494/4321)

These exercises illustrate the importance of thinking "conditionally" by showing that the overall pattern of data can be reversed when looking at different groups separately (that is, conditional on groups). And this is why we need to use a multiple regression model when we have more than one predictor at play. Suppose we set up a model that predicts the success rates in our climbing example. Two predictors we have are i) Climbers A and B and ii) route grades. If we run two simple regressions on i) and ii) separately, we will end up with two unconditional results that B performs better than A from i) and Grade V8 has a lower success rate than V4 from ii), the former of which is misleading as we discussed above. A multiple regression model avoids this issue because it includes both predictors in the same regression and the effect of each predictor is estimated conditional on the others.

Technically, the group variable in the Simpson's paradox can be understood as an example of a *confounder*, a variable that messes up with a causal link between other variables, if omitted. And, there are many interesting examples of confounders in statistical analysis. We will discuss this point in more detail in Topic 7 since it plays a very important role in statistical causal inference.

1.2 Hidden causation

Hidden causation is a statistical phenomenon where the influence of a predictor on the outcome is felt through another predictor. For example, suppose there is some positive relationship in data between a mother's age at the time she gives birth and a child's test score at age of 3. Does this mean that we should recommend that parents try to have a baby as late as possible? Well, there are several mechanisms by which a mother's age affects a child's test score. For example, a mother's age at her childbirth could be related to her educational level, and what matters to her child's test score might not be her age but her educational level. Or, it might be the case that older mothers who gave birth to a child at a later age in the data might already have another kid so that they know how to raise a kid better. In all these cases, there are hidden links of causality

behind the impact of a mother's age on her child's test score.

A multiple regression model can help us find out this hidden causation. In a mother's age vs. child test score case, we can first try a simple regression model only with a mother's age as a predictor and check the coefficient. Then, we repeat this regression that further includes a mother's education level at the time of birth and check if the coefficient of the predictor for a mother's age has changed. If it has significantly decreased while the coefficient of a mother's education level is positive, this indicates the possibility that there is a hidden causal link behind the impact of a mother's age on her child's test score.

1.3 Interaction

Another good reason for using a multiple regression model is the potential interactions between predictors. This is the case when two or more predictors have a non-additive effect on the outcome variable such as a synergy effect. For example, it is often argued in cancer studies that chemotherapy and radiation therapy are interacting variables in cancer treatment. That is, the effectiveness of chemotherapy increases when treated with radiation therapy and vice versa. In this case, the standalone simple linear regression either overstate or understate the effectiveness of each chemotherapy and radiation therapy without capturing the synergy effect of two. A multiple regression model with an additional interacting term can address this deficiency of a simple linear regression as will be shown below.

2 Basic framework

Multiple linear regression is a statistical technique that uses more than one predictor to predict the outcome variable. Assume that we have a sample of n observations on outcome variable y and $k-1$ explanatory variables, $x_1, \ldots, x_{k-1}$. We choose $k-1$ to make the total k number of coefficients including the intercept. Then, a multiple regression model is written as follows[1]

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1} + \varepsilon \\ &= \beta_0 + \sum_{j}^{k-1} \beta_j x_j + \varepsilon \end{aligned} \tag{1}$$

This is an extension of a simple linear regression model by adding more predictors on the right-hand side. The only notational difference is that we use β_0 instead of α for the intercept.[2] As before, we keep the same error structure by assuming that errors are iid and are normally distributed with mean zero and a constant variance $\varepsilon \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$.

Not surprisingly, the OLS, MLE, and Bayesian estimators (with uninformative priors) for the intercept and slope coefficients can be derived exactly the same way as we did in the simple regression case. Since the estimators of coefficients derived from these three different methods are the same, we will illustrate the multiple linear regression using the OLS framework and focus more on the interpretation of the coefficients and the actual examples of the multiple linear regression in Section 3.

[1] x_j, y, and ε are all $n \times 1$ vectors.

[2] This is because we want to use a matrix notation for the OLS derivation, and denoting the intercept by β_0 makes it notationally convenient to collect all the coefficients in β vector.

As before, the OLS estimators for β_0, $\beta_1, \ldots, \beta_k$ can be derived by solving the following optimization problem:

$$\arg\min_{\boldsymbol{\beta}} \sum_{i}^{n} (y_i - \beta_0 - \sum_{j}^{k-1} \beta_j x_{ij})^2 \tag{2}$$

The solution to this optimization problem is extremely messy in non-matrix form. For example, $\hat{\beta}_1$ for the two explanatory variable case, which is the simplest multiple regression, is the following:

$$\hat{\beta}_1 = \frac{(\sum_i^n x_{i2}^2)(\sum_i^n x_{i1} y_i) - (\sum_i^n x_{i1} x_{2i})(\sum_i^n x_{i2} y_i)}{(\sum_i^n x_{i1}^2)(\sum_i^n x_{i2}^2) - (\sum_i^n x_{i1} x_{i2})^2}$$

This shows that the estimation of each of the coefficients requires product and cross summation of all variables so the notation will become unwieldy when we add more explanatory variables. Therefore, it is standard to use a matrix notation for multiple regression analysis.

Using a matrix form, we can write multiple regression as

$$\mathbf{y} = \mathrm{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3}$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathrm{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k-1} \\ 1 & x_{21} & \cdots & x_{2k-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk-1} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{k-1} \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

$\mathbf{y}$ is a $n \times 1$ vector, X is a $n \times k$ matrix, $\boldsymbol{\beta}$ is a $k \times 1$ vector, and ε is a $n \times 1$ vector. The OLS estimator for $\hat{\boldsymbol{\beta}}$ can be derived by solving the following optimization problem:[3]

$$\arg\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathrm{X}\,\boldsymbol{\beta})'(\mathbf{y} - \mathrm{X}\,\boldsymbol{\beta}) \tag{4}$$

whose solution is

$$\hat{\boldsymbol{\beta}} = (\mathrm{X}'\,\mathrm{X})^{-1}\,\mathrm{X}'\,y \tag{5}$$

The expected value and the variance of the estimator in a matrix form are

$$\mathrm{E}(\hat{\boldsymbol{\beta}}|\,\mathrm{X}) = \boldsymbol{\beta} \tag{6}$$

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}|\,\mathrm{X}) = \sigma^2(\mathrm{X}'\,\mathrm{X})^{-1} \tag{7}$$

Same as the simple regression, it can be shown that the sampling distribution for $\boldsymbol{\beta}$ is a multivariate normal distribution with the mean and variance from Eqs. 6 and 7.

$$\hat{\boldsymbol{\beta}} \sim \mathrm{N}\left(\boldsymbol{\beta},\ \sigma^2(\mathrm{X}'\,\mathrm{X})^{-1}\right) \tag{8}$$

[3]See Appendix for proof.

When σ^2 is unknown, its estimator is[4]

$$\hat{\sigma^2} = \frac{(\mathbf{y} - \mathrm{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathrm{X}\hat{\boldsymbol{\beta}})}{n-k} = \frac{1}{n-k}\sum_{i}^{n}(y_i - \hat{y}_i)^2 \tag{9}$$

As we discussed in Topic 4, when σ is unknown with small sample size, the student-t distribution is a correct sampling distribution for $\hat{\boldsymbol{\beta}}$

$$\hat{\boldsymbol{\beta}} \;\sim\; \mathrm{t}_{n-k}\left(\boldsymbol{\beta},\ \sigma^2(\mathrm{X}'\mathrm{X})^{-1}\right). \tag{10}$$

3 Examples and interpretations of multiple regression model

Now that we set up a basic estimation framework for multiple regression, we will turn our attention to examples and interpretations of the multiple regression model.

3.1 Multiple predictors with no interactions

Changes of intercepts

Suppose we have two predictors with no interactions (two predictors which do not interact in their influence on y) with the following specification

$$y \;=\; \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{11}$$

Here, we have two predictors x_1 and x_2 with a normal error term ε with zero mean. To get some intuition about multiple regression models, let us first assume that x_1 is a continuous variable and x_2 is a discrete variable taking the value of 0 or 1. We can think of x_2 as a group variable (group 0 and 1). Table 1 shows how the data is structured.

Table 1: Sample data

Obs.	y	x_1	x_2(group)
1	-2.50	-1.20	0
2	0.53	1.89	1
3	4.34	0.11	0
4	-0.76	-0.15	1
⋮	⋮	⋮	
198	-4.11	-0.29	1
199	0.59	-0.89	0
200	7.14	2.65	0

Figure 1 visualizes the data in two different ways. The left panel shows x_1 versus y with a regression line from a simple linear regression of y on x_1 (without x_2). The right panel shows the same relationship conditional on the group variable x_2. A simple regression line is added for each group, that is, the regression line obtain from individual estimation of y over x1 for each group. It is clear from the figure that the slope of the regression line without the group variable x_2 (black) is quite similar to that of the regression lines for each group (blue and red). However, the simple linear regression without group predictors (left) fails to capture the crucial pattern that the

[4]See Topic 4 for an intuitive explanation of why we need $n-k$ correction for σ^2. A formal proof can be found in the Appendix.

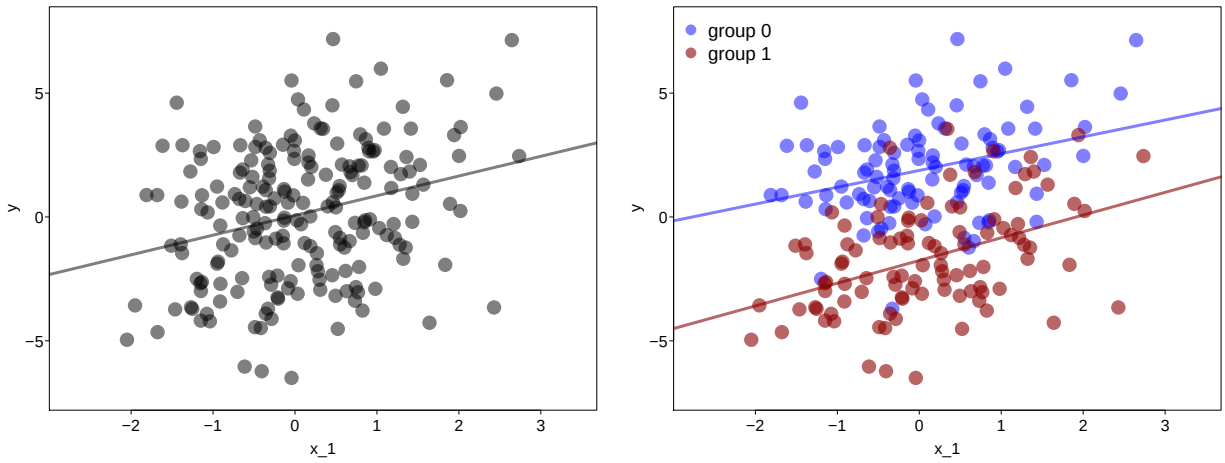


Figure 1: x vs. y, pooled & group-level

observations in group 1 systematically have smaller y values than those in group 0. As shown in the blue and red regression lines, the intercept of the group 1 line is significantly lower than that of group 0.

To understand how multiple regression models can capture the difference in intercepts across groups, let's now look at Table 2 which compares two regression results with and without the group predictor x_2. The value in the parenthesis is one standard deviation of the sampling distribution for each coefficient. In the simple regression without the group variable x_2, the slope coefficient of x_1 (β_1) and the intercept (β_0) are 0.795 and 0.062, respectively. The interpretation is straightforward: β_0 is the y-intercept on a $x - y$ plane (the value of y is 0.062 when $x_1 = 0$) while β_1 is the marginal impact of x_1 on y (0.792 increase of y when x increases by 1).

Table 2: Regression regression with and without x_2

	Outcome variable:	
	y	
	Model without x_2	Model with x_2
β_1	0.795	0.808
	(0.192)	(0.138)
β_2		-3.615
		(0.263)
β_0	0.062	1.869
	(0.184)	(0.187)
# of Obs.	200	200
R^2	0.080	0.530

In the multiple regression, the coefficients of x_1 and x_2 (β_1 and β_2), and the intercept β_0 are,

0.808, -3.615, and 1.869, respectively:

$$y = 1.869 + 0.808x_1 - 3.615x_2 + \varepsilon$$

The interpretation of these coefficients is not as straightforward as the simple regression case. First, the intercept β_0 is the value of y when *both* x_1 and x_2 are zero. In our example, it coincides with the y intercept for Group 0 (both $x_1 = 0$ and $x_2 = 0$), that is, the intercept of the blue line in Figure 1.[5]

Second, the coefficient of x_1 (β_1) is the *partial effect* of x_1 on y, meaning that it captures the relationship between y and x_1 when x_2 is fixed (that is, no matter what the values of x_2 are). More intuitively, we can think of the value of $\beta_1 = 0.808$, as the best approximation to the slope of a regression line that works for both of our two different groups. It doesn't work as perfectly as the individual regression lines in the right panel of Figure 1 in predicting the slope for each specific group but is the best approximation when two groups are considered simultaneously.

Finally, β_2 is the partial effect of x_2 on y (given x_1 constant). In our exercise, it reflects the difference in y value between Group 0 and 1: the y values in group 0 smaller than those in Group 0 by 3.615. We can think of this as a change in the intercept of the two separate regression lines in the right panel of Figure 1. From Equation 12, we see that, for Group 0, the intercept is $\beta_0 = 1.869$ (when $x_1 = 0$ and $x_2 = 0$) and for Group 1, the intercept is $\beta_0 - 3.615 = -1.746$ (when $x_1 = 0$ and $x_2 = 1$).

Simpson's paradox

Let's repeat the same multiple linear specifications of Equation 11 with data. In this example, x_1 is a discrete variable indicating 0 = Male and 1 = Female and x_2 is a group variable with 0, 1, indicating two different departments.[6] Figure 2 visualizes data with regression lines. Contrary to Figure 1, the sign of the slope of the simple regression line (left plot) reverses when we look at the regression line of each of the individual groups (right plot). While x_1 and y have a positive relationship in either of two groups, it becomes negative when we ignore groups.

As we discussed in Section 1, this seemingly puzzling phenomenon is called *Simpson's paradox* where including an additional predictor reverses or nullifies the association between existing predictors and the outcome variable. In the context of college admission, the main predictor x_1 is the gender indicator and the group predictor x_2 is the department indicator. y is the admission score. When looking at the overall relationship between gender and the admission score from simple linear regression, there is a negative relationship, meaning that female applicants tend to have a lower score compared to their male counterparts. This relationship is misleading: when we look at the same relationship for each department (Group 0 and 1), the relationship is reversed and now the female candidates tend to have higher admission scores.

A multiple regression model can address this deficiency of the simple linear regression model. Table 3 compares two regression results with and without the group predictor x_2. As expected, the slope coefficient in the model without x_2 is negative (-8.281). In contrast, the model with an additional group predictor estimated the partial effect of gender

[5] If predictors cannot be zero, the intercept has no intrinsic meaning.

[6] It is worthwhile to note that ANOVA and OLS with categorical predictors are mathematically identical even though the model output takes a different form. This course does not directly discuss ANOVA modeling and replace it with detailed discussions on multiple and multilevel regression modeling.

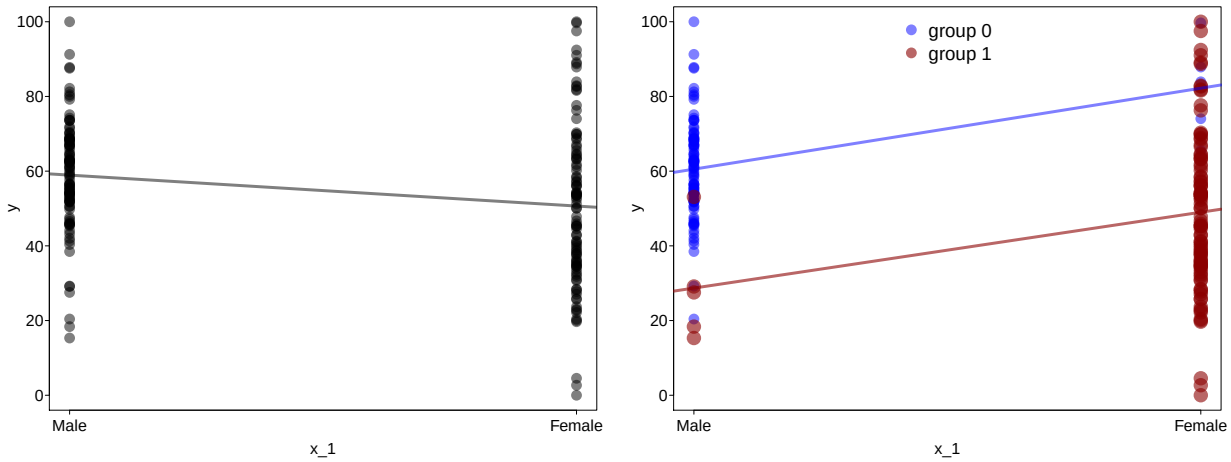


Figure 2: Multiple regression example: Simpson's paradox

on admission scores, β_1 to be positive with a significant magnitude, 20.992. β_2 is the difference in the average admission score between two departments, reflecting the changes in the intercept in blue and red regression lines in Figure 2. Group 0 has higher admission scores than Group 1 by 32.525. Finally, β_0 is the admission score for Group 0 ($x_2 = 0$) when $x_1 = 0$.

Table 3: Regression result for Simpson's paradox example

	Outcome variable:	
	y	
	Model without x_2	Model with x_2
β_1	-8.281	20.992
	(2.645)	(5.625)
β_2		-32.525
		(5.625)
β_0	58.938	60.564
	(1.870)	(1.756)
Observations	200	200
R^2	0.047	0.185

The reason behind the change in the sign of the impact of gender on admission score is because substantially more female candidates applied to the department where the admission score tends to be lower. Therefore, even though female candidates outperform male candidates in both departments, the higher presence of female candidates in a tough department makes their performance look bad when we look at the overall score.

Hidden causation

As a final example of a multiple regression model without interaction, Let's visit the mother's age vs. child test score case we briefly discussed in Section 1 with actual data from Gelman & Hill (2006). The dataset has three columns as shown in Table 4: Child's test score at age 3, Mother's

educational level, and Mother's age.

Table 4: Data on child test score from Gelman & Hill (2006).

Obs.	Child's Score	Mother's Educ. Level	Mother's Age
1	120	2	21
2	89	1	17
3	78	2	19
4	42	1	20
⋮	⋮	⋮	
399	98	1	18
400	81	2	22

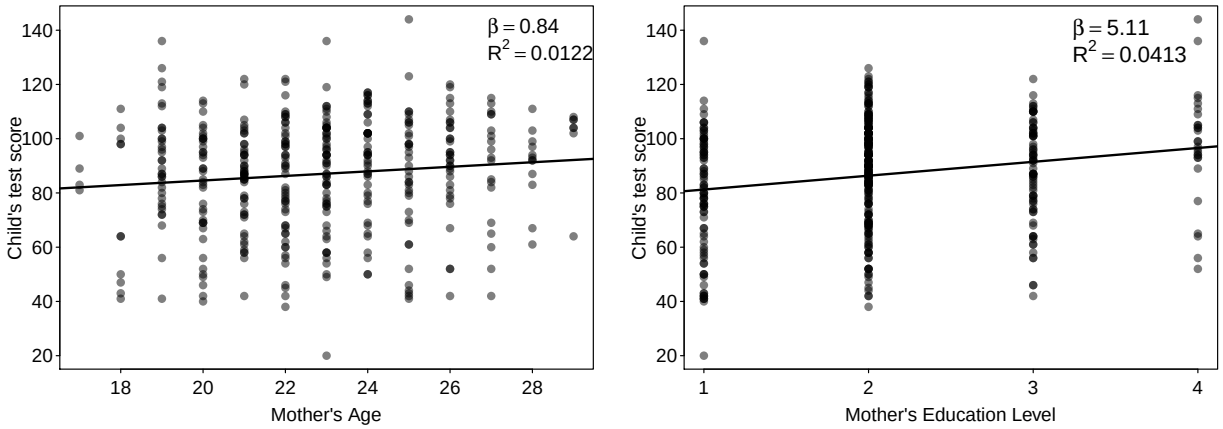


Figure 3: Child's test score vs. Mother's age and Mother's educational level.

First, we are interested in the impact of a mother's age on a child's test score at age 3. The left plane in Figure 3 visualizes the regression result, showing that the older a mother is, the higher her child's score with the slope of 0.84. Should we then recommend that mothers give birth as late as possible? Let's be more careful and use an additional piece of information from the data set, namely, a mother's educational level (1-3 scales). The right panel visualizes the regression result and also shows that there is a positive relationship between a mother's educational level and a child's test score. Then what variable should we trust more in predicting a child's test score?

As we discussed in Section 1, comparing simple and multiple regression models can help us to find out how which predictor plays the key role or not by tracking the changes in the magnitude of the coefficients. Table 5 compares three regression results: a model only with x_1, a model only with x_2, and a model both with x_1 and x_2.

As the result shows, the impact of a mother's age reduces dramatically from 0.840 to 0.343 when we run a multiple regression model with an additional predictor of a mother's education level. The standard error of this coefficient in the multiple regression is so high (0.398) that we cannot be sure if the true impact is above zero. In contrast, the impact of a mother's education level changes very little and has a similar value both in the simple and multiple regression models.

Table 5: Multiple regression for Child test score example

	Outcome variable:		
	Child's test score		
	Model with x_1	Model with x_2	Model with x_1 and x_2
β_1 (Mother age)	0.840		0.343
	(0.379)		(0.398)
β_2 of x_2 (Mother Edu level)		5.107	4.711
		(1.233)	(1.317)
β_0	67.783	76.143	69.155
	(8.688)	(2.792)	(8.571)
Observations	400	400	400
R^2	0.012	0.041	0.043

R^2 does not change either. A mechanism behind this result goes like this. A mother's age in itself does not have any predictive power for a child's test score. The reason we see a positive relationship between these two variables in the simple regression is because the impact of a mother's education level is felt through a mother's age. When we include a mother's education level and estimate the impact of a mother's age, we see a dramatic decrease in the magnitude and the predictive power of this predictor.

To understand why a mother's age loses a predictive power when adding a mother's educational level as another predictor, Figure 4 visualizes the impact of a mother's age on a child's score conditional on a mother's educational level. The sign of the regression line flips in 4 different education levels, explaining why we have reduced magnitude and predictive power of a mother's age when we include a mother's educational level as another predictor.

3.2 Multiple predictors with interactions

Now let us turn our attention to another intriguing example that highlights the strength of the multiple linear regression model. Here, we will introduce *interactions* between predictors. Equation 12 shows the regression specification with two predictors and their interaction.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon \tag{12}$$

The key difference between the regression models with and without interactions is that the impact of the predictor comes from two different sources. β_1, for example, still represents the impact of predictor x_1 but it is not the unique effect. When there is an interaction, the impact also comes from the interaction coefficient β_3 whose magnitude depends on the values of interacting predictor, x_2.[7]

To get some intuition, let us illustrate this point with data. As before, x_1 is a continuous variable and x_2 is a discrete variable taking the value of 0 and 1. Figure 5 visualizes the linear

[7]Taking a partial derivative of y with respect to x_1 makes it clear that the partial effect of x_1 depends on x_2: $\frac{\partial y}{\partial x_1} = \beta_1 + \beta_3 x_2$

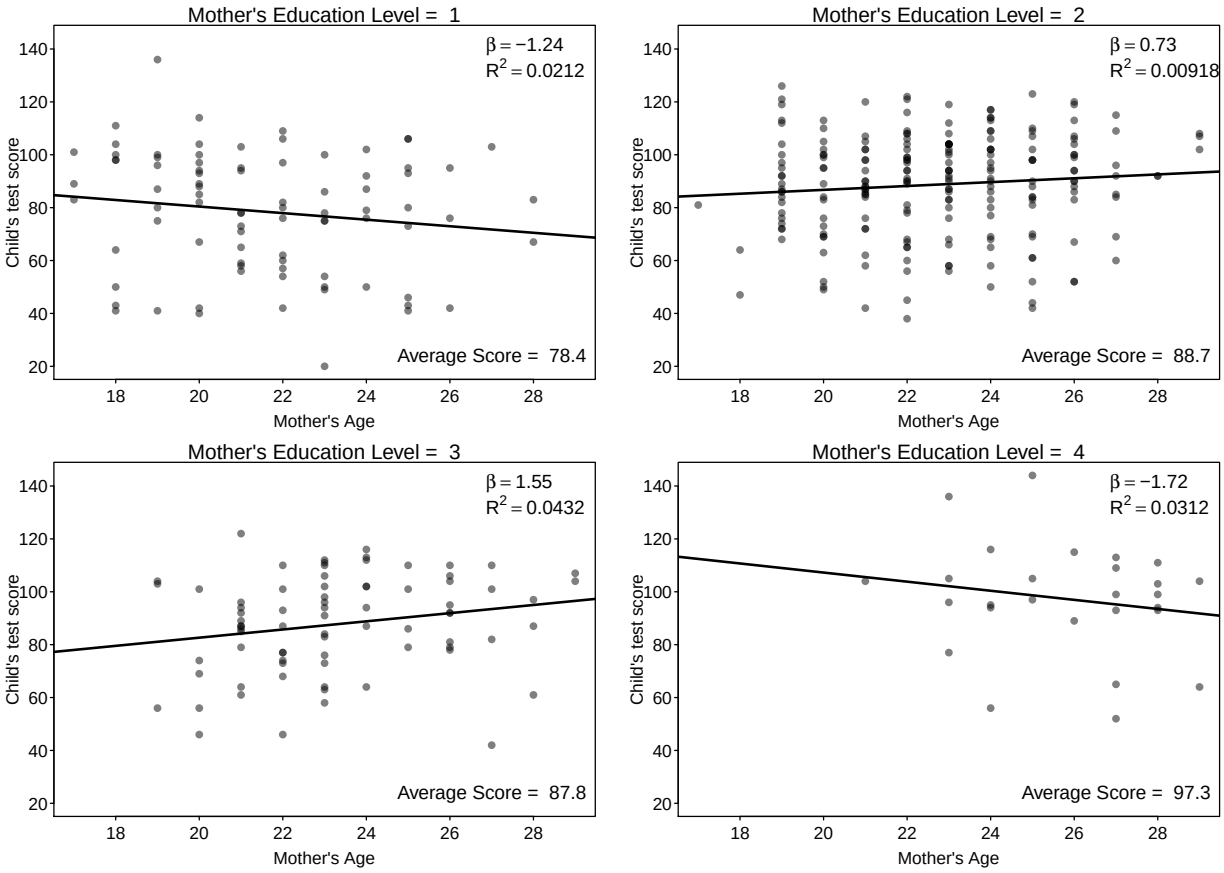


Figure 4: Child's test score vs. Mother's age conditional on Mother's educational level.

relationship between y and x_1 (left) and the same relationship conditional on x_2 (right). The key difference between Figure 1 and Figure 5 is that the slopes of two individual regression lines for each group have an opposite sign. This small change makes a dramatic difference in the regression model with and without interactions.

To see this point, Table 6 compares three different regression results: a model only with x_1, a model both with x_1 and x_2, and a model both with x_1 and x_2 and their interaction $x_1 \times x_2$. The first simple linear regression model fails to capture the completely different pattern of a linear relationship in two groups. Interestingly, the second regression model both with x_1 and x_2 fails either and does not capture the positive slope in Group 1 as can be shown in the negative coefficient of β_1 (-0.197). β_2 in this model only determines the changes in the $y-$intercept and does not capture the changes in the slope of lines between two groups.

Contrast to the first two models, the model with an interaction correctly captures the changes in the slope across groups. Let's see how we can tease out this information from the results:

$$y = 1.818 - 1.482x_1 + 0.548x_2 + 2.357x_1x_2 + \varepsilon$$

First, as before, $\beta_0 = 1.818$ is the $y-$intercept for Group 0 ($x_1 = 0$ and $x_2 = 0$). For Group 1, the $y-$intercept is $1.818 + 0.548 = 2.366$ ($x_1 = 0$ and $x_2 = 1$). The partial effect of x_1 comes from two different sources: the standalone coefficient $\beta_1 = -1.482$ and the

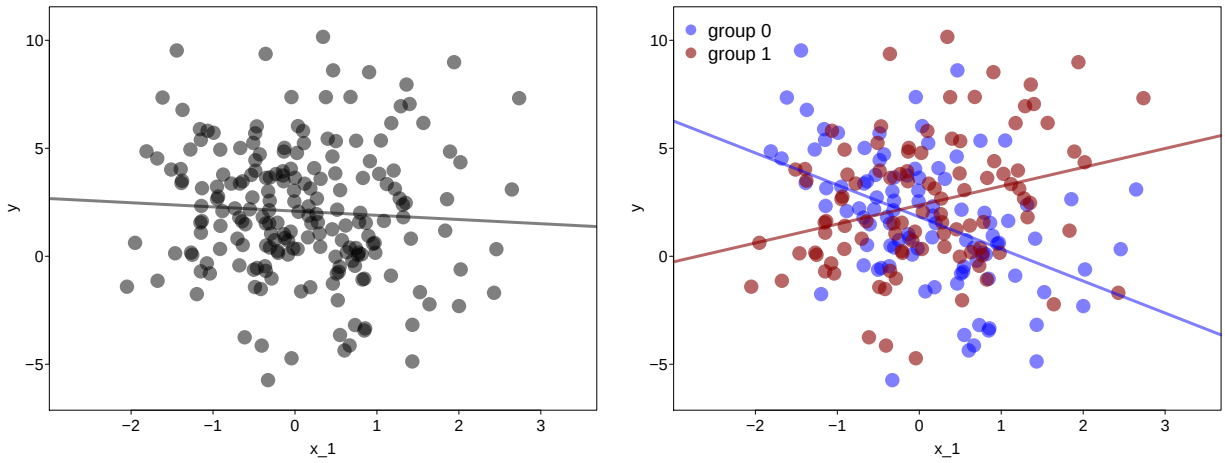


Figure 5: Multiple regression example: varying slope

interaction coefficient $\beta_3 = 2.357$. Most importantly, the interaction coefficient depends on which group we are looking at. For Group 0, β_3 is multiplied by 0, so the partial effect of x_1 is $-1.482 + (2.3570) = -1.482$. So, β_1 is interpreted as the unique effect of x_1 on y only for Group 0 ($x_2 = 0$). For Group 1, β_3 is multiplied by 1, so the partial effect of x_1 is $-1.482 + 1 \times 2.357 = 0.875$. That is, $\beta_1 + \beta_3$ is interpreted as the unique effect of x_1 on y only for Group 1 ($x_2 = 1$). This result correctly represents the changes in the slope of regression lines in the right panel of Figure 5: -1.482 for Group 0 in blue and 0.875 for Group 1 in red.

This result suggests that when we have multiple predictors and want to understand their impact on a certain outcome variable, it is the best practice to include their interaction terms as much as possible to account for potential changes in the linear relationship across groups[8].

[8]This applies to continuous variables as well since we can interpret a continuous variable as containing infinitely many groups.

Table 6: Regression result for models with interactions

	Outcome variable:		
	y		
	Model only with x_1	Model with x_1 and x_2	Model with x_1, x_2 and $x_1 \times x_2$
β_1	-0.194	-0.197	-1.482
	(0.224)	(0.222)	(0.306)
β_2		0.742	0.548
		(0.425)	(0.397)
β_3 of $x_1 \times x_2$			2.357
			(0.415)
β_0	2.090	1.719	1.818
	(0.215)	(0.301)	(0.280)
Observations	200	200	200
R^2	0.004	0.019	0.158

A OLS estimator in matrix form

Consider the linear regression model[9]

$$\mathbf{y} = \mathrm{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

$\mathbf{y}$ is a $n \times 1$ vector, X is a $n \times k$ matrix, $\boldsymbol{\beta}$ is a $k \times 1$ vector, and ε is a $n \times 1$ vector. Assume that $\boldsymbol{\varepsilon} \overset{\text{iid}}{\sim} \mathrm{N}(\mathbf{0}, \sigma_\epsilon^2 I)$.

The OLS minimizes the sum of squared residuals, which we denote by L:

$$\begin{aligned} \arg\min_{\boldsymbol{\beta}} \mathrm{L} &= (\mathbf{y} - \mathrm{X}\,\boldsymbol{\beta})'(\mathbf{y} - \mathrm{X}\,\boldsymbol{\beta}) \\ &= \mathbf{y}'\mathbf{y} - \boldsymbol{\beta}'\,\mathrm{X}'\,\mathbf{y} - \mathbf{y}'\,\mathrm{X}\,\boldsymbol{\beta} + \boldsymbol{\beta}'\,\mathrm{X}'\,\mathrm{X}\,\boldsymbol{\beta} \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\,\mathrm{X}\,\boldsymbol{\beta} + \boldsymbol{\beta}'\,\mathrm{X}'\,\mathrm{X}\,\boldsymbol{\beta} \end{aligned} \tag{13}$$

The necessary condition for a minimum is:

$$\frac{\partial\,\mathrm{L}}{\partial\boldsymbol{\beta}} = -2\,\mathrm{X}'\,\mathbf{y} + 2\,\mathrm{X}'\,\mathrm{X}\,\boldsymbol{\beta} = 0 \tag{14}$$

Let $\hat{\boldsymbol{\beta}}$ be the solution. Then, $\boldsymbol{\beta}$ satisfies $\mathrm{X}'\,\mathrm{X}\,\boldsymbol{\beta} = \mathrm{X}'\,\mathbf{y}$ which results from the necessary condition above. If the inverse of $\mathrm{X}'\,\mathrm{X}$ exists, which it does by assumption (assumption of full rank), then the solution is:

$$\hat{\boldsymbol{\beta}} = (\mathrm{X}'\,\mathrm{X})^{-1}\,\mathrm{X}'\,\mathbf{y} \tag{15}$$

which is the OLS estimator.

[9] See Greene (2018) for more detailed proof.

The expected value of $\boldsymbol{\beta}$ is derived as follows. Since $\hat{\boldsymbol{\beta}} = (\mathrm{X}'\,\mathrm{X})^{-1}\,\mathrm{X}'\,\mathbf{y}$, we can write

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathrm{X}'\,\mathrm{X})^{-1}\,\mathrm{X}'(\mathrm{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= (\mathrm{X}'\,\mathrm{X})^{-1}\,\mathrm{X}'\,\mathrm{X}\,\boldsymbol{\beta} + (\mathrm{X}'\,\mathrm{X})^{-1}\,\mathrm{X}'\,\boldsymbol{\varepsilon} \\ &= \boldsymbol{\beta} + (\mathrm{X}'\,\mathrm{X})^{-1}\,\mathrm{X}'\,\boldsymbol{\varepsilon} \end{aligned} \tag{16}$$

Therefore,

$$\begin{aligned}\mathrm{E}(\hat{\boldsymbol{\beta}}|\,\mathrm{X}) &= \boldsymbol{\beta} + \mathrm{E}((\mathrm{X}'\,\mathrm{X})^{-1}\,\mathrm{X}'\,\boldsymbol{\varepsilon}|\,\mathrm{X}) \\ &= \boldsymbol{\beta} + (\mathrm{X}'\,\mathrm{X})^{-1}\,\mathrm{X}'\,E(\boldsymbol{\varepsilon}|\,\mathrm{X}) \\ &= \boldsymbol{\beta} \end{aligned} \tag{17}$$

since $\mathrm{E}(\boldsymbol{\varepsilon}|\,\mathrm{X})$ is assumed to be zero.

The variance of the error vector $\boldsymbol{\varepsilon}$ is derived as follows. We have $\hat{\boldsymbol{\beta}} = (\mathrm{X}'\,\mathrm{X})^{-1}\,\mathrm{X}'(\mathrm{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\mathrm{X}'\,\mathrm{X})^{-1}\,\mathrm{X}'\,\boldsymbol{\varepsilon}$. So, the variance of the least squares estimator is:

$$\begin{aligned}Var(\hat{\boldsymbol{\beta}}|\,\mathrm{X}) &= \mathrm{E}((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'|\,\mathrm{X}) \\ &= \mathrm{E}((\mathrm{X}'\,\mathrm{X})^{-1}\,\mathrm{X}'\,\varepsilon\varepsilon'\,\mathrm{X}(\mathrm{X}'\,\mathrm{X})^{-1}|\,\mathrm{X}) \\ &= (\mathrm{X}'\,\mathrm{X})^{-1}\,\mathrm{X}'\,\mathrm{E}(\varepsilon\varepsilon'|\,\mathrm{X})\,\mathrm{X}(\mathrm{X}'\,\mathrm{X})^{-1} \\ &= (\mathrm{X}'\,\mathrm{X})^{-1}\,\mathrm{X}'(\sigma^2 I)\,\mathrm{X}(\mathrm{X}'\,\mathrm{X})^{-1} \\ &= \sigma^2(\mathrm{X}'\,\mathrm{X})^{-1} \end{aligned} \tag{18}$$

where the last step follows from the assumption $\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma_{\varepsilon}^2 I$.

The proof of $\hat{\sigma^2} = \frac{1}{n-k}\sum_i^n (y_i - \hat{y}_i)^2$ is mathematically more involved. Let's first define the vector of least squares residuals $\mathbf{e}$, which we will call residuals as before,

$$\begin{aligned}\mathbf{e} &= \mathbf{y} - \mathrm{X}\,\hat{\boldsymbol{\beta}} \\ &= \mathbf{y} - \mathrm{X}(\mathrm{X}'\,\mathrm{X})^{-1}\,\mathrm{X}'\,\mathbf{y} \\ &= (\mathrm{I} - \mathrm{X}(\mathrm{X}'\,\mathrm{X})^{-1}\,\mathrm{X}')\mathbf{y} = \mathrm{M}\,\mathbf{y} \end{aligned} \tag{19}$$

where the $n \times n$ matrix M is the "residual maker," the matrix that produces the residuals when it pre-multiplies any vector $\mathbf{y}$. Note that $\mathrm{MX} = 0$, because if X is regressed on X, we will have a perfect fit and the residuals will be zero. This makes

$$\begin{aligned}\mathbf{e} &= \mathrm{M}\,\mathbf{y} = \mathrm{M}(\mathrm{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= \mathrm{M}\,\boldsymbol{\varepsilon} \end{aligned} \tag{20}$$

since $\mathrm{M}\,\mathrm{X}\,\boldsymbol{\beta} = 0$. From here, we can calculate the expected value of the sum of squared residuals as an (unnormalized) estimator of σ^2. The expected value of the sum of squared residuals is

$$\mathrm{E}(\mathbf{e}'\mathbf{e}|\,\mathrm{X}) = \mathrm{E}(\boldsymbol{\varepsilon}'\,\mathrm{M}\,\boldsymbol{\varepsilon}|\,\mathrm{X}) \tag{21}$$

Note that $\mathbf{e}'\mathbf{e} = \boldsymbol{\varepsilon}'\,\mathrm{M}'\,\mathrm{M}\,\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}'\,\mathrm{M}\,\boldsymbol{\varepsilon}$ since the residual maker M is idempotent, meaning that its multiplication yields itself: $\mathrm{M}\,\mathrm{M} = \mathrm{M}'\,\mathrm{M} = \mathrm{M}\,\mathrm{M}' = \mathrm{M}$.

We can rewrite $\boldsymbol{\varepsilon}' \, \mathrm{M} \, \boldsymbol{\varepsilon}$ in terms of its trace, the sum of elements on the main diagonal of the matrix, since it is a scalar (1×1 matrix). Using the property of the trace $\mathrm{tr}(\mathrm{CAC}) = \mathrm{tr}(\mathrm{ACC})$, we obtain

$$\mathrm{E}(\mathrm{tr}(\boldsymbol{\varepsilon}' \, \mathrm{M} \, \boldsymbol{\varepsilon}) | \, \mathrm{X}]) = \mathrm{E}(\mathrm{tr}(\mathrm{M} \, \boldsymbol{\varepsilon}' \boldsymbol{\varepsilon}) | \, \mathrm{X}]) \tag{22}$$

which can be rewritten as

$$\mathrm{tr}(\mathrm{ME}(\varepsilon' \boldsymbol{\varepsilon} | \, \mathrm{X})) = \mathrm{tr}(\mathrm{M} \, \sigma^2 \, \mathrm{I}) = \sigma^2 \, \mathrm{tr}(\mathrm{M}) \tag{23}$$

Here, an important result is derived from the trace of M. Since $\mathrm{M} = \mathrm{I_n} - \mathrm{X}(\mathrm{X}' \, \mathrm{X})^{-1} \, \mathrm{X}'$, its trace is $\mathrm{tr}(\mathrm{I_n}) - \mathrm{tr}((\mathrm{X}' \, \mathrm{X})^{-1} \, \mathrm{X}' \, \mathrm{X})$. Note that $-\mathrm{tr}((\mathrm{X}' \, \mathrm{X})^{-1} \, \mathrm{X}' \, \mathrm{X}) = \mathrm{tr}(\mathrm{I_K})$ due to the similarity invariance of the trace that $\mathrm{tr}\left(\mathrm{P}^{-1} \, \mathrm{AP}\right) = \mathrm{tr}(\mathrm{A})$, where in our case A is an identity matrix I. Therefore,

$$\mathrm{tr}(\mathrm{M}) = \mathrm{I_n} - \mathrm{I_K} = n - K. \tag{24}$$

Then we have

$$\mathrm{E}(\mathbf{e}'\mathbf{e} | \, \mathrm{X}) = (n - K)\sigma^2 \tag{25}$$

which leads to our estimator for σ^2

$$\begin{aligned} \hat{\sigma}^2 &= \frac{\mathbf{e}'\mathbf{e}}{n - K} \\ &= \frac{1}{n - k} \sum_{i}^{n} (y_i - \hat{y}_i)^2 \end{aligned} \tag{26}$$

References

Gelman, A. & Hill, J. (2006), *Data analysis using regression and multilevel/hierarchical models*, Cambridge university press.

Greene, W. (2018), *Econometric Analysis. 8th edition*, Pearson.

Chapter 6: Generalized linear regression model

Jangho Yang

v1.4

Contents

1 Issues with the linear regression model

Even though the linear regression model is a powerful tool for analyzing correlation in data, its scope is rather limited due to the normality assumption of errors. For example, when the outcome variable is discrete, we cannot use the linear regression with normal errors because a normal distribution is defined for continuous variables. The same goes for a bounded outcome variable. When we know that the outcome variable is bounded between 0 and 1 , it does not make too much sense to assume that the error of the outcome is distributed normally. Generalized linear regression models we will discuss in this section address this deficiency of the linear regression model with normal errors and consequently allow us to analyze relations in many different types of variables.

2 Basic framework

To avoid the use of the matrix algebra, the mathematical exposition of the model will only involve a simple linear prediction with one predictor.

2.1 Link function

The linear regression with a normal error can be expressed as follows:

$$y \sim \mathrm{N}(\mu, \sigma^2) \tag{1}$$
$$\mu = \alpha + \beta x \tag{2}$$

This expression says that the outcome variable y is normally distributed with a linear mean prediction of $\alpha + \beta x$. The generalized linear model (GLM) generalizes the normality assumption but keep the linear prediction part.

$$y \sim f(\theta, \phi) \tag{3}$$
$$g(\theta) = \alpha + \beta x \tag{4}$$

f is any outcome distribution with two sets of parameters: θ and ϕ. θ is related to the linear predictor through g: $g(\theta) = \alpha + \beta x$. This function g is called a *link function* and it links the linear prediction $\alpha + \beta x$ to the parameter of the outcome distribution f. It can take any form depending on the type of the parameter of the outcome distribution. ϕ represents the additional set of parameters that are not part of linear prediction but are necessary for the outcome distribution to work. The GLM is a "generalized" version of the linear regression because we can freely choose f and g depending on types of data and our hypotheses.

To better understand how f and g can be chosen, let's look into two widely-used GLMs: *Binomial-logistic regression* and *Poisson regression.*

2.2 Binomial-logistic regression: outcome variable either 0 or 1

In real life, we encounter all sorts of binary outcomes such as pass/fail, hit/miss, life/death, or win/lose. Suppose we have some variables to predict the chance of pass or fail, e.g. the duration of study time. How can we construct a regression model in this case? A *Binomial-logistic regression model* is a tool for examining the correlation between this binary outcome and some linear predictors. Before we set up f and g for this regression model, let's look at a type of

data we are dealing with. Table 1 shows two example datasets with the discrete outcome (y) and one predictors (x): binary outcomes (left) and repeated binary outcomes out of n trials (right).

Obs.	y	n	x
1	0	1	8.45
2	1	1	9.02
3	1	1	10.56
4	1	1	11.35
⋮	⋮	⋮	⋮
98	1	1	7.35
99	0	1	8.14
100	1	1	10.36

Obs.	y	n	x
1	3	6	9.70
2	2	8	8.00
3	3	7	9.79
4	2	6	7.41
⋮	⋮	⋮	⋮
48	4	5	8.60
49	3	7	10.93
50	3	5	12.81

Table 1: Data with (repeated) binary outcomes y

The key characteristic of the data is that the outcome variable y is either 0 or 1 or the sum of repeated binary outcomes given n. To model this, we need to specify an outcome distribution that gives either 0 or 1 (or the sum of repeated binary outcomes given n). The *Binomial-logistic regression model* uses the binomial distribution as the outcome distribution f since its random variable is the sum of the repeated binary outcomes y given the success probability p and the number of trials n. When $n = 1$, the binomial distribution is equivalent to the Bernoulli distribution whose random variable takes the value 1 with probability p and the value 0 with probability $1 - p$. In this model, the linear predictor $\alpha + \beta x$ determines the probability of success p given the link function g:

$$\begin{aligned} y &\sim \text{Binomial}(n, p) \\ g(p) &= \alpha + \beta x \end{aligned} \tag{5}$$

That is, the outcome variable is distributed according to the binomial distribution and the probability parameter as a function of the linear predictor. Then, what is a right link function g that links the linear predictor $\alpha + \beta x$ to the probability p? Here, it is more intuitive to think of this problem by taking the inverse of the link function, that is $p = g^{-1}(\alpha + \beta x)$. In our binomial outcome distribution, p is a "probability" so the output of g^{-1} should lie between zero and one. There are several candidates for such a function, such as a CDF of any distributions (but particularly a normal distribution) but we will focus on a particular function called *logistic function* in this section. This is because the logistic function is easy to interpret and is mathematically more tractable compared to its counterparts. The logistic function is defined as follows

$$\text{logistic}(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}, \tag{6}$$

where e^x is the exponential function. The logistic function looks like an S-shaped curve and converges to 0 and 1 (see Figure 1), meaning that it properly links the domain of $\alpha + \beta x$ to the probability domain of p between 0 and 1.

Using this logistic function as the "inverse" link function $g^{-1}(.) = \text{logistic}(.)$, we have $\text{logistic}^{-1}(p) = \alpha + \beta x$. This inverse logistic function is known as the *logit* function

$$\text{logistic}^{-1}(p) = \text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta x \tag{7}$$

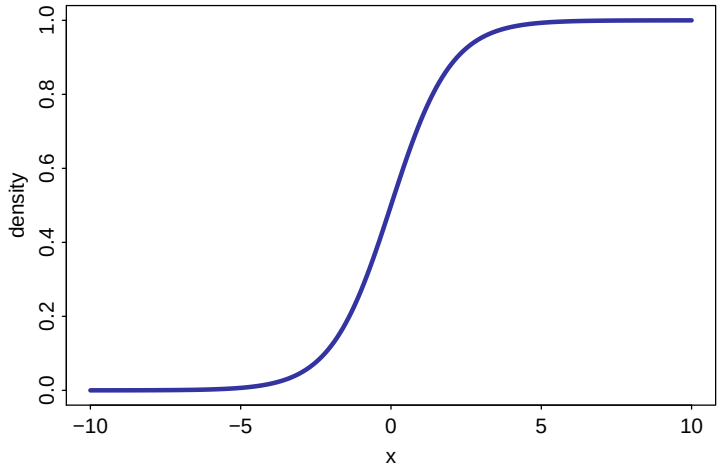


Figure 1: Logistic distribution.

Note that the logit function is often called the log-odds function. The odds gives how more likely the probability (of success) as opposed to the probability of failure. For example, if the success probability $p = 0.9$, then the odds is 9, meaning that the success is 9 times more likely than the failure. The log-odds is just the logarithm of the odds and changes its scale from $[0, \infty]$ to $[-\infty, \infty]$. Despite its mathematical usefulness, the log-odds is not always intuitive to interpret. Therefore, we will use the logit function as a mathematical too to link p and $\alpha + \beta x$ without paying too much attention to log odds interpretation. Using the logit link function, we now have a complete model specification for binomial-logistic regression.

$$\begin{aligned} y &\sim \text{Binomial}(n, p) \\ \text{logit}(p) &= \alpha + \beta x \end{aligned} \tag{8}$$

We will examine how to estimate and interpret this binomial-logistic regression in Sections 3 and 4.

2.3 Poisson regression: integer outcome variable

Let's now introduce one more GLM, called *Poisson regression*. This model is useful when the outcome variable is an integer (count data), e.g. the number of patients waiting when you arrive at the hospital between 2 and 3 pm, the number of bombs hitting in the south of London during World War II. Unlike the binomial outcome variables, it does not have to be based on a number of independent trials.

The outcome variable y is modeled using the Poisson distribution with the rate parameter λ whose functional form is

$$f(y; \lambda)) = \frac{\lambda^y e^{-\lambda}}{y!}. \tag{9}$$

Both the expected value and the variance are $\text{E}[y] = \text{Var}(y) = \lambda$. See Figure 2.

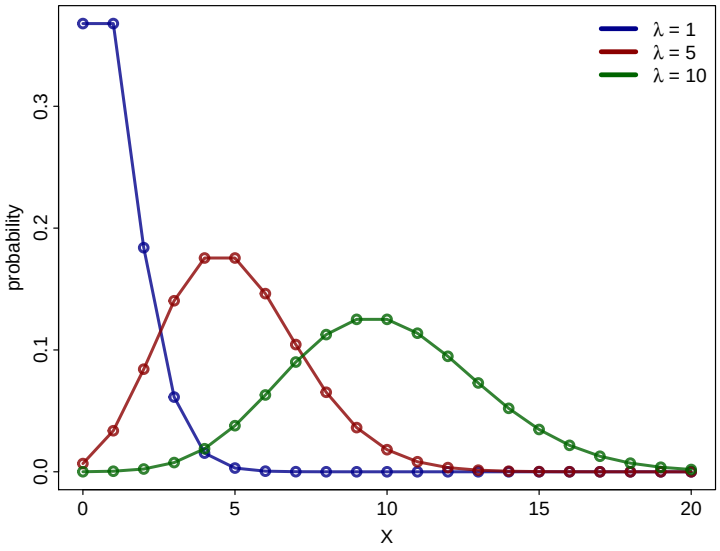


Figure 2: Poisson distribution with varying parameter λ.

The parameter $\lambda > 0$ represents the rate parameter or the average number of occurrence. In the Poisson regression, the linear predictors $\alpha + \beta x$ determine this rate parameter given the link function g:

$$\begin{aligned} y &\sim \text{Poisson}(\lambda) \\ g(\lambda) &= \alpha + \beta x \end{aligned}$$

Then, what is the right link function g for Poisson regression? Again, let's first take the inverse of the link function: $\lambda = g^{-1}(\alpha + \beta x)$. We can immediately see that the output of g^{-1} should be greater than zero. For this matter, the Poisson regression uses an exponential function since its output is always positive. Using this exponential function as the "inverse" link function $g^{-1}(x) = \mathrm{e}^x$, we have $\ln(\lambda) = \alpha + \beta x$. Therefore, the complete model specification of Poisson regression is

$$y \sim \text{Poisson}(\lambda) \tag{10}$$
$$\ln(\lambda) = \alpha + \beta x \tag{11}$$

For both binomial-logistic and Poisson regression, the relationship between the linear prediction and the outcome variable is mediated by the link function. This makes the interpretation of the coefficients more challenging as we will discuss in Section 4.

3 Model estimation

We will now discuss how to estimate these two GLMs. It is important to note that GLMs violate the normality assumption of errors so we cannot use the OLS method. For this reason, the GLMs are often estimated by MLE or Bayesian method. In this section, we will briefly discuss the basic framework of the MLE estimation of logistic and Poisson regression.

3.1 Binomial-logistic regression

As we discussed in Topic 3, the MLE finds the hypothesis $\hat{\theta}$ that has the maximum likelihood value

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \widehat{L}_m(\theta \,|y) \tag{12}$$

where θ is a set of parameters of the model and m is the number of observations. The functional form of the binomial distribution with the parameter p is $f(y;p) = \binom{n}{y} p^y (1-p)^{n-y} \quad \text{for } y \in \{0,1\}$ meaning that the probability of success is p and the probability of failure $1-p$. We can drop the binomial coefficient $\binom{n}{y}$ since it is a constant. The likelihood function is then $\widehat{L}_m(p\,|y) = \prod_{i=1}^{m} p_i^{y_i}(1-p_i)^{n_i - y_i}$ and the log-likelihood of the binomial-logistic regression can be written as

$$\begin{aligned}
\widehat{l}_m(p|y) &= \ln \prod_{i=1}^{m} p_i^{y_i}(1-p_i)^{n_i - y_i} \\
&= \sum_{i=1}^{m} y_i \ln(p_i) + \sum_{i=1}^{m} (n_i - y_i)\ln(1-p_i) \\
&= \sum_{i=1}^{m} n_i \ln(1-p_i) + \sum_{i=1}^{m} y_i \ln\left(\frac{p_i}{1-p_i}\right)
\end{aligned}$$

Since $(1-p_i) = \frac{1}{e^{\alpha+\beta x_i}+1}$ and $\ln(\frac{p_i}{1-p_i}) = \alpha + \beta x_i$ from Equation 7, we can rewrite the log-likelihood as follows

$$\widehat{l}_m(p|y) = \sum_{i=1}^{m} n_i \ln(e^{\alpha+\beta x_i}+1)^{-1} + \sum_{i=1}^{m} y_i(\alpha + \beta x_i) \tag{13}$$

Therefore, to get the MLE estimator for α, β, we need to solve the following optimization problem

$$\hat{\alpha}, \hat{\beta} = \arg\max_{\alpha,\beta} \sum_{i=1}^{m} n_i \ln(e^{\alpha+\beta x_i}+1)^{-1} + \sum_{i=1}^{m} y_i(\alpha + \beta x_i) \tag{14}$$

However, it is not possible to solve this program exactly and find a closed-form expression for the estimator. This is because when we set the derivatives of the log-likelihood equal to zero, we have α, β appearing both as an argument to an exponential function and a linear addition. This type of equation, which is called *transcendental equations* is not solvable analytically. (See 4.1 for an example) This means that we need to use some approximation methods. Luckily, there are well-established algorithms that enable us to find the MLE estimators approximately. Examining these algorithms in detail is beyond the scope of this course. Instead, we will briefly discuss some intuitions behind the numerical approximation method in Section 4.

3.2 Poisson regression

Now let's set up the MLE for Poisson regression.The functional form of the Poisson distribution is $f(y;\lambda) = \dfrac{\lambda^y e^{-\lambda}}{y!}$. Therefore, the likelihood function is $\widehat{L}_n(\lambda\,|y) = \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!}$ and the log-likelihood

of the Poisson regression can be written as

$$\begin{aligned}\widehat{l}_n(\lambda|y) &= \ln\prod_{i=1}^{n}\frac{\lambda^{y_i}e^{-\lambda}}{y_i!}\\ &= \sum_{i=1}^{n}\ln\left(\frac{e^{-\lambda}\lambda^{y_i}}{y_i!}\right)\\ &= -\sum_{i=1}^{n}\lambda+\sum_{i=1}^{n}y_i\ln(\lambda)-\sum_{i=1}^{n}\ln(y_i!). \end{aligned} \tag{15}$$

Since $\lambda = \mathrm{e}^{\alpha+\beta x}$ from Equation 11, we can rewrite the log-likelihood as follows

$$\widehat{l}_n(\lambda|y) = -\sum_{i=1}^{n}e^{\alpha+\beta x_i}+\sum_{i=1}^{n}y_i(\alpha+\beta x_i)-\sum_{i=1}^{n}\ln(y_i!). \tag{16}$$

As we did with the binomial-logistic regression, we need to solve the following optimization problem to get the MLE estimator for α, β

$$\hat{\alpha},\hat{\beta} = \arg\max_{\alpha,\beta}\ -\sum_{i=1}^{n}e^{\alpha+\beta x_i}+\sum_{i=1}^{n}y_i(\alpha+\beta x_i)-\sum_{i=1}^{n}\ln(y_i!). \tag{17}$$

Same as the binomial-logistic regression, there is no closed-form solution to this problem due to the transcendental equation. We will show below how to use an approximation method to get α, β.

4 Examples and interpretation of estimated parameters

4.1 Examples of binomial-logistic regression

Data and MLE estimator

Suppose we have 4 data points of y which are modelled to be binomially distributed:

$$y_i \sim \text{Binomial}(n_i, p_i),$$

where p_i is the probability of success given a predictor x_i. The data are

Obs.	y	n	x
1	1	9	0.71
2	0	1	0.25
3	2	5	0.39
4	3	7	0.09

The linear predictor is expressed through the logit link function is $\text{logit}(p_i) = \alpha+\beta x_i$. Therefore, $\text{logit}^{-1}(\alpha+\beta x_i) = \frac{e^{(\alpha+\beta x_i)}}{1+e^{(\alpha+\beta x_i)}}$, and our log likelihood function in terms of the parameters α and β is:

$$
\begin{aligned}
\widehat{l}_4(\alpha, \beta | y, x, n) &= \sum_i^4 y_i \log\left(\text{logit}^{-1}(\alpha + \beta x_i)\right) + \sum_i^4 (n_i - y_i) \log\left(1 - \text{logit}^{-1}(\alpha + \beta x_i)\right) \\
&= \sum_i^4 y_i \log\left(\frac{e^{(\alpha+\beta x_i)}}{1 + e^{(\alpha+\beta x_i)}}\right) + \sum_i^4 (n_i - y_i) \log\left(1 - \frac{e^{(\alpha+\beta x_i)}}{1 + e^{(\alpha+\beta x_i)}}\right) \\
&= \sum_i^4 y_i \left(\alpha + \beta x_i - \log(1 + e^{(\alpha+\beta x_i)})\right) - (n_i - y_i) \left(\log(1 + e^{(\alpha+\beta x_i)})\right) \\
&= \sum_i^4 y_i(\alpha + \beta x_i) - n_i \log(1 + e^{(\alpha+\beta x_i)})
\end{aligned}
$$

Now, let's get the FOC (the analytical gradient) of the log likelihood function:

$$
\begin{aligned}
\frac{\partial \widehat{l}_4(\alpha, \beta | y, x, n)}{\partial \alpha} &= \sum_{i=1}^4 y_i - \frac{n_i e^{(\alpha+\beta x_i)}}{1 + e^{(\alpha+\beta x_i)}} \\
\frac{\partial \widehat{l}_4(\alpha, \beta | y, x, n)}{\partial \beta} &= \sum_{i=1}^4 y_i x_i - \frac{n_i x_i e^{(\alpha+\beta x_i)}}{1 + e^{(\alpha+\beta x_i)}}
\end{aligned}
$$

When we set this derivative to zero, we have $\alpha + \beta x_i$ appearing both as an argument to an exponential function. As we discussed in Section 3, this transcendental equation is not solvable analytically. Instead, we need a numerical approximation.

Numerical approximation

The most primitive approach is to directly plug some possible numbers of α and β and pick the pair that has the highest log-likelihood. This is called a *grid approach*. To get some intuition, Table 2 shows the normalized log-likelihood for some values of α and β. It shows that α roughly between -0.25 and 0.25 and β between -3 and -2 has the highest log-likelihood value.

Table 2: Normalized log-likelihood for some values of α (rows) and β (columns)

	-5	-4.5	-4	-3.5	-3	-2.5	-2	-1.5	-1	-0.5	0
-1	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01
-0.75	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.02	0.02	0.01	0.01
-0.5	0.00	0.00	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.01	0.01
-0.25	0.00	0.01	0.01	0.01	0.02	0.02	**0.03**	0.02	0.01	0.01	0.00
0	0.01	0.01	0.01	0.02	**0.03**	**0.03**	0.02	0.02	0.01	0.00	0.00
0.25	0.01	0.01	0.02	0.02	**0.03**	0.02	0.02	0.01	0.00	0.00	0.00
0.5	0.01	0.02	0.02	0.02	0.02	0.01	0.01	0.00	0.00	0.00	0.00
0.75	0.01	0.02	0.02	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00
1	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00

We can use more fine-grained α and β values to get a more accurate picture of the high log-likelihood region. Figure 3 visualizes the log-likelihood given some possible ranges of α and β.

It is clear that some areas have higher log-likelihood (yellow) while other areas have lower values (red). If we use a more fine-grained sequence of α and β, we can get a more clear picture. This primitive grid approximation is the basis of the standard numerical optimization algorithms,

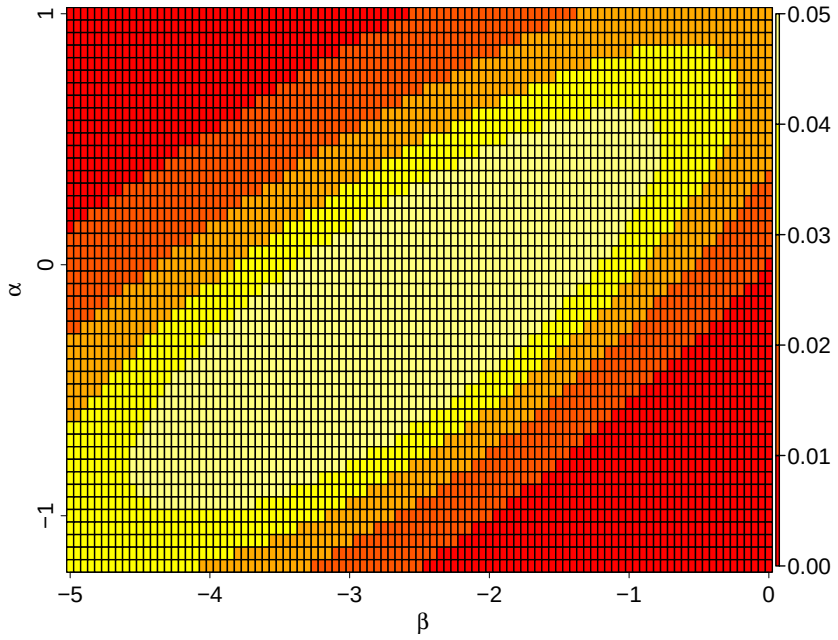


Figure 3: Visualization of the log likelihood given some rages of α and β.

which in effect use the numerical gradient of the objective function (log-likelihood function) given the very small change (similar to the fine-grained sequence of the grid) of the parameters. When the small change, which we denote by ϵ, is very small enough, the difference between the analytical and the numerical gradient becomes negligible. We can see this by plugging some random numbers in both analytical and numerical gradient and calculate the difference.

The numerical gradient of the log likelihood can be obtained by finding its rate of change given very small change of α and β:

$$\frac{\Delta \widehat{l}_4(\alpha, \beta|y, x, n)}{\Delta \alpha} = \frac{\widehat{l}_4(\alpha + \epsilon, \beta \mid y, x, n) - \widehat{l}_4(\alpha - \epsilon, \beta \mid y, x, n)}{2\epsilon}$$
$$\frac{\Delta \widehat{l}_4(\alpha, \beta|y, x, n)}{\Delta \beta} = \frac{\widehat{l}_4(\alpha, \beta + \epsilon \mid y, x, n) - \widehat{l}_4(\alpha, \beta - \epsilon \mid y, x, n)}{2\epsilon}$$

We calculate the difference between the analytical and numerical gradient by plugging 10,000 pairs of random numbers generated from Uniform$(-10, 10)$. The mean of the difference is

$$\frac{\partial \widehat{l}_4(\alpha, \beta|y, x, n)}{\partial \alpha} - \frac{\Delta \widehat{l}_4(\alpha, \beta|y, x, n)}{\Delta \alpha} = 0.000000157$$
$$\frac{\partial \widehat{l}_4(\alpha, \beta|y, x, n)}{\partial \beta} - \frac{\Delta \widehat{l}_4(\alpha, \beta|y, x, n)}{\Delta \beta} = -0.0000000259$$

which are almost zero, implying that the analytical and numerical gradients have almost the same value. The following is the histogram of each difference, again showing that the distributions degenerate very close to zero. This result gives some confidence about the reliability of the

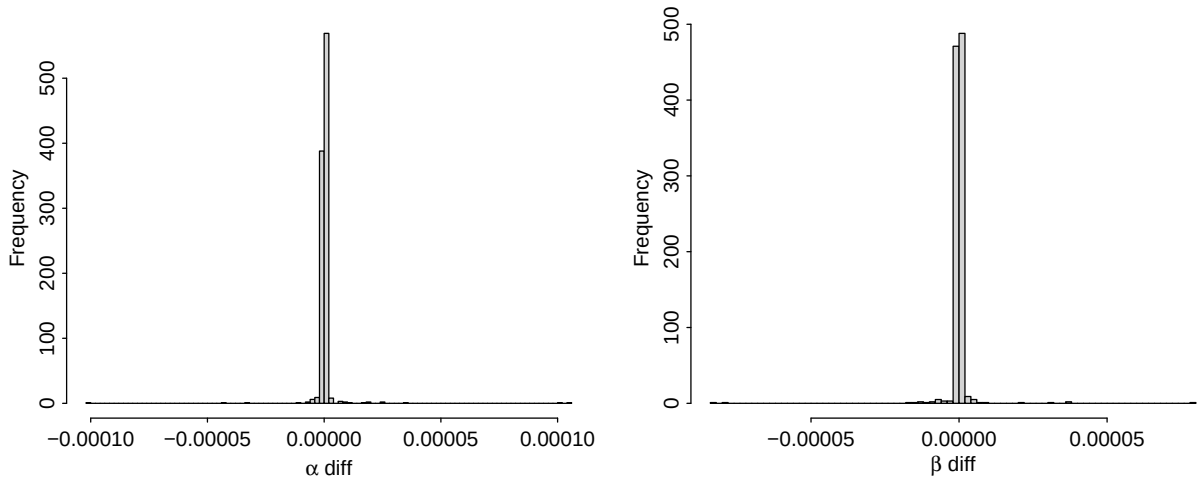


Figure 4: Visualization of the log likelihood given some ranges of α and β.

numerical approximation of MLE estimators (at least for simple problems).

Numerical optimization (or computational algorithm in general) is the backbone of the statistical work and has gained more and more attention these days as the models become more complex and the size of data becomes larger. Discussing further details of numerical optimization is beyond the scope of this course. We will instead use some standard numerical optimization techniques (Newton-type algorithms) to solve our GLM models.

Different statistical softwares have different optimization tools. For R, `nlm` and `optim` functions are most widely used.[1] Using `nlm`, we can numerically approximate the MLE estimators as follows:

$$\begin{aligned} \hat{\alpha} &= -0.00163 \\ \hat{\beta} &= -2.587 \end{aligned}$$

which is roughly the center of the grid in Figure 3.

Interpretation of coefficients

Then, how can we interpret these coefficients in the link function?

$$p = \frac{1}{1 + e^{-(-0.00163 - 2.587x)}}$$

Before we answer this question, let's get some intuition by drawing our logistic link function 8 with varying α and β values in Figure 5. The left plane shows that α shifts the location of the logistic curve. Same as the linear regression, α needs to be evaluated assuming $x = 0$. When $\alpha = 0$, the logistic curve centers at 0, meaning that the probability of $x = 0$ is 0.5. When $\alpha = 3$ the curve (blue) shifts left and has a higher probability at any point of x compared to the curve with $\alpha = 0$. At $x = 0$, the curve with $\alpha = 3$ has $p = 0.952$.[2] In contrast, when $\alpha = -3$, the curve (green) shifts right and has a lower probability at any point of x compared to the curve with

[1] `glm` function is specially designed for GLM models but is less flexible.

[2] The curve shifts from zero by $-\alpha/\beta$ because $\alpha + \beta * x = \beta(x + \alpha/\beta)$.

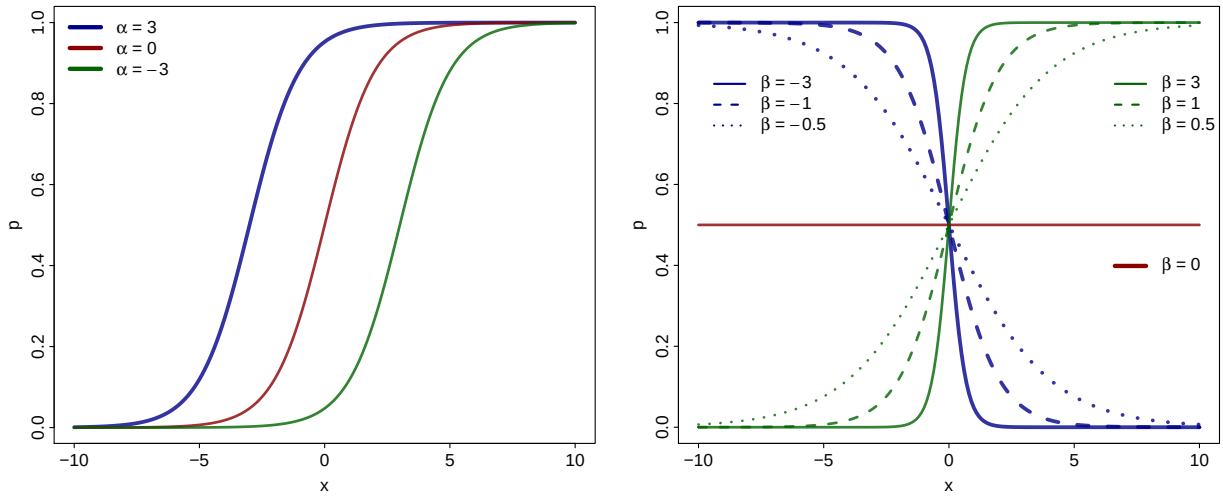


Figure 5: Logistic link function with varying values of α and β.

$\alpha = 0$. At $x = 0$, the curve with $\alpha = -3$ has $p = 0.047$.[3]

The right plane shows that β determines the shape of the curve. Most obviously, when it is negative, it has a negative relationship between p and x, and makes p a monotonically decreasing function of x. When positive, p and x have a positive relationship. Further, β determines the overall curvature of the curve. The larger $|\beta|$, the steeper the curvature. Note that the actual slope (the first derivative) of the logistic curve depends on x. The first derivative of the logistic curve $\text{logit}^{-1}(\alpha + \beta x_i) = 1/(1 + e^{-(\alpha+\beta x_i)})$ is $\beta e^{-(\alpha+\beta x_i)}/(1 + e^{-(\alpha+\beta x_i)})^2$. This slope can be evaluated at any points of interest, but is often evaluated at the central value since the slope is steepest at its center and gives its upper bound. If it is evaulated at the center of the logistic curve $-\alpha/\beta$ (see footnote 2), the slope always is $4/\beta$. This is the divide by 4 rule.

Since α determines the location of the curve, $\alpha = -0.00163$ in our exercise, meaning that when $x = 0$, $p = 0.4995$.

$\beta = -2.587$ suggests that x and p have a negative relationship. At $x = 0$, we have $-2.587e^{-0.00163}/(1 + e^{-0.00163})^2 = -0.647$. This means that a unit increase in x leads to -0.647 reduction in p when the reference point is at $x = 0$.

4.2 Examples of Poisson regression

Data and MLE estimator

Suppose we have a pair of discrete data as shown in Table 3. We can visualize x vs. y in Figure 6.

Since the outcome variable y is discrete, we will be using a Poisson model as follows

$$y_i \sim \text{Poisson}(\lambda),$$

[3]When the predictor cannot be zero, the intercept loses intrinsic meaning and needs to be evaluated at some fixed value of x, e.g. the mean of the predictor.

Table 3: Poisson Data

Obs.	y	x
1	3	1
2	5	2
3	2	3
⋮	⋮	⋮
8	37	8
9	35	9
10	51	10

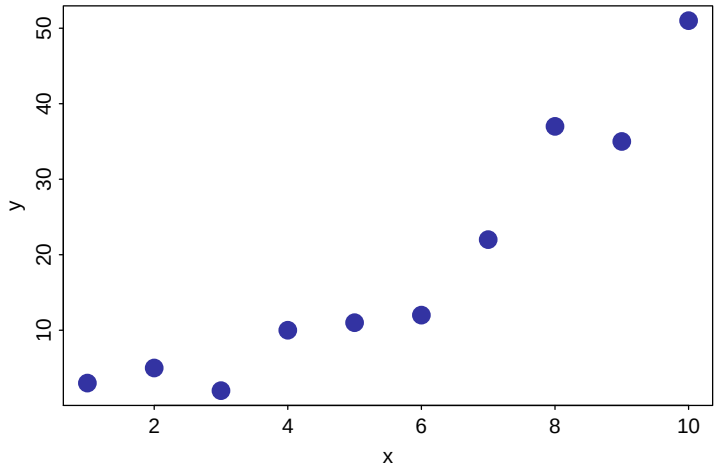


Figure 6: Visualization of data in Table 3.

where λ is the rate parameter of the Poisson distribution. The linear predictor is expressed through the exponential link function as follows:

$$\log(\lambda) = \alpha + \beta x_i.$$

Therefore, our log-likelihood function in terms of the parameters α and β is:

$$\begin{aligned} \hat{l}(\alpha, \beta | y, x) &= \log(\prod_{i=1}^{10} \frac{(e^{(\alpha+\beta*x_i)})^{y_i} e^{(-e^{(\alpha+\beta x_i)})}}{y_i!}) \\ &= \sum_{i=1}^{10} \log(\frac{(e^{(\alpha+\beta*x_i)})^{y_i} e^{(-e^{(\alpha+\beta x_i)})}}{y_i!}) \\ &= \sum_{i=1}^{10} y_i(\alpha + \beta x_i) - e^{(\alpha+\beta*x_i)} - \log(y_i!) \end{aligned} \tag{18}$$

The first derivative are:

$$\begin{aligned} \frac{\partial log(p(\alpha, \beta|y))}{\partial \alpha} &= \sum_{i=1}^{10} y_i - e^{(\alpha+\beta*x_i)} \\ \frac{\partial log(p(\alpha, \beta|y))}{\partial \beta} &= \sum_{i=1}^{10} y_i x_i - x_i e^{(\alpha+\beta x_i)} \end{aligned}$$

Similar to the binomial-logistical MLE, when we set this derivative to zero, we have $\alpha + \beta x_i$ appearing as an argument to an exponential function, and cannot be solved analytically. Instead, we need a numerical approximation.

Numerical approximation

Using `nlm`, we can find the MLE estimators as follows:

$$\begin{aligned} \hat{\alpha} &= 0.786 \\ \hat{\beta} &= 0.320 \end{aligned}$$

Interpretation of coefficients

How can we interpret these coefficients in the link function?

$$\lambda = e^{(0.786+0.320x)}$$

$\alpha = 0.786$ is the intercept of the regression with $x = 0$. This means that the rate parameter λ is $e^{(0.786)} = 2.195$ at $x = 0$, leading to the expected value of y being 2.195 at $x = 0$ due to the property of the Poisson distribution that $E[X] = \lambda$. The coefficient β is the expected change in y *on the logarithmic scale* (or the percentage change) for a unit change in x. This is because a change in x is linked through the exponential function. To understand this point, let's take look at Figure 7.

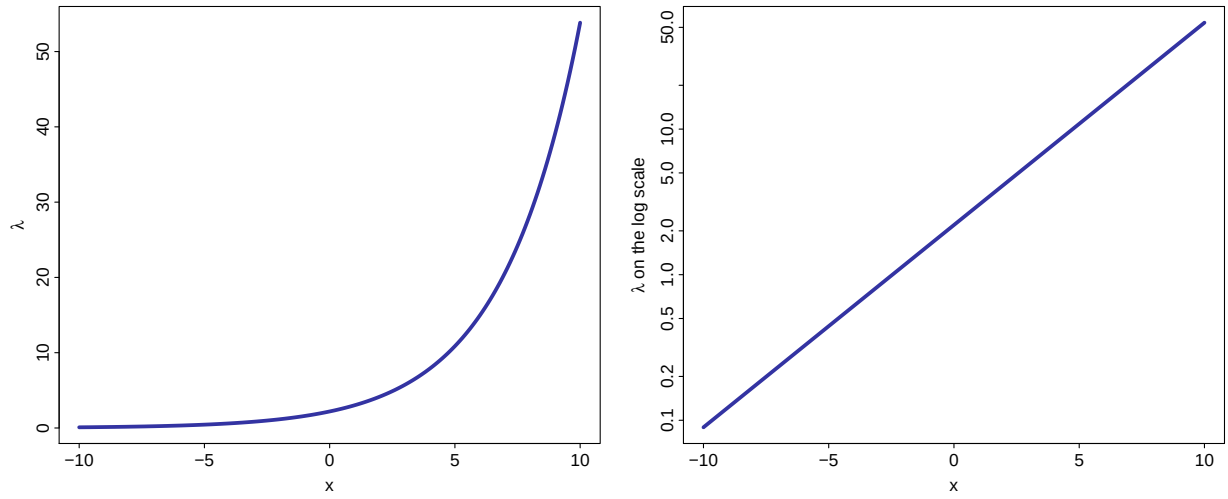


Figure 7: Exponential functions on the non-log and log scale

The left panel is the exponential function $e^{(0.786+0.320x)}$ on the non-log scale and the right panel is the exact same exponential function on the log scale of y. It shows that the exponential function is log-linear, meaning that a unit increase in x leads to a unit increase in λ on the log scale.

In our example, therefore, $\beta = 0.320$ means that a unit increase in x predicts $(e^{0.32} - 1) \times 100\%$ increase in the expected value of y.

Chapter 7: Multilevel/Hierarchical Linear Regression

Jangho Yang

v1.0

Contents

Topic 6 discussed the limitations of the linear regression model and introduced generalized linear models that allow for a more flexible choice of outcome variables. Topic 7 will show another way of generalizing the linear regression model and discuss the hierarchical linear model, also known as the varying coefficient, multilevel, or mixed effect model. The very basic idea of the hierarchical linear model is that the regression coefficients vary by different groups but their variations are structured by a higher level determinant, often called *hyperparameters*. This model is useful when there are heterogeneous groups in the data and we want to understand variation across groups without overfitting the model. Due to the complex nature of the hierarchical model, it gained popularity only recently with advanced computing power.

Learning basics about the hierarchical linear model is not too challenging. The concepts are straightforward, the mathematical representation is intuitive, and the model implementation is reasonably easy thanks to programming packages. One key impediment, however, lies in semantics. Some of the key concepts in the hierarchical model, such as fixed/random effects, have been used in widely different contexts across different disciplines. For this reason, some students, especially those who have learned these concepts one way or another, often experience a steep learning curve in the course. To eliminate the source of confusion, this chapter will discuss how these basic concepts are understood in different statistical models and help students to focus more on the substance of the model than on semantics.

1 Multigroup data

1.1 Data structure

The type of data we are interested in is *multigroup* data in which observations can be batched into multiple groups. This grouping is justified by the heterogeneous characteristics inherent in each group (e.g. gender, religion, country). Figure 1 shows an example of multigroup data: the firm-level financial statement, which has group-variables such as country, year, and sector.

As can be seen in this example, the multigroup data can include a time variable, meaning that there are repeated observations of the same variables (e.g. company's sales) over some periods of time. When data includes measurements over time along with other group-level variables (e.g. country or sector), it is often called *panel data*. There is a reason why the time dimension is distinguished from other group-level variables: time-series measurements are often correlated. For example, if the GDP in Canada grows faster than average in one year, it is likely that it will grow faster than average as well in the following year. Therefore, when the data involves time-series measurement, we should consider the possibility of *autocorrelation*, a cross-unit dependence of variables across time. A detailed discussion on autocorrelation in panel data, however, is beyond the scope of this course. Therefore, this chapter will focus on the regression model using particular multigroup data with no cross-unit dependence over time.

1.2 Potential issues

Even without the issue of autocorrelation, analyzing multigroup data poses unique challenges that can not be easily solved within a simple regression framework. These challenges result from potential heterogeneity across different groups. Further, we often do not know what causes the group-level heterogeneity due to the lack of data that could account for the source and the degree

Figure 1: Multilevel data example: Corporate financial statement

of heterogeneity.

This unique feature of multigroup data raises two important research questions. i) if the linear relationship between variables behaves differently across groups, how can we obtain the overall linear relationship between variables across different groups while adjusting for group-level heterogeneity? ii) if we are interested in group-level difference itself, how can we set-up a single regression model such that we can estimate the degree of group-level difference in target coefficients (intercepts and slopes) without repeating the same regression for different groups?

In the following sections, we will overview several regression models to answer these questions.

1.3 Simpson's paradox revisited

To better understand some of the issues involving multigroup data, let's revisit Simpson's paradox, a statistical phenomenon where including another predictor reverses or nullifies the association between existing predictors and the outcome variable. As we discussed in Topic 5, this concept has a close connection to regression with multigroup data since including group-variable can have an impact on the target coefficients. To see this point, suppose we are interested in the impact of speeding fines on car accidents in three different countries, A, B, and C.[1] Table1 shows the average number of car accidents and the average fine for speeding tickets. It shows that the correlation between these two variables is positive, meaning that when the average speeding ticket is more costly, there are more

[1]We will be ruling out reverse causality for the sake of simplicity.

car accidents. This is somewhat counter-intuitive since a higher penalty should reduce car accidents.

Table 1: Car accident vs. speeding ticket

Country	# of Accidents	Ticket fine ($)
A	229	100
B	239	103
C	250	105

Now, let's look at the full panel data in Table 2. The correlation between the number of car accidents and speeding tickets becomes negative when we look at each of the counties, A, B, and C, separately.

Table 2: Car accident vs. speeding ticket, full panel

Country	Year	# of Accidents	Ticket fine ($)
A	2000	221	102
	2001	223	102
	2002	232	100
	2003	233	100
	2004	237	98
	2005	231	99
B	2000	243	101
	2001	242	103
	2002	237	104
	2003	231	106
	2004	244	100
	2005	234	104
C	2000	251	104
	2001	249	106
	2002	250	106
	2003	254	103
	2004	245	107
	2005	249	105

This point can be seen more clearly in Figure 2. The left side plot shows the linear relationship between car accidents and speeding tickets without the country-level variable, meaning that observations are *pooled* across all three countries. Similar to the positive correlation found in Table 1, there is a clear positive relationship. In contrast, the right side plot shows the same relationship conditional on the country-group. Here, the sign of the linear relationship changes, and the car accidents and speeding tickets have a negative relationship. This is typical of Simpson's paradox situation and suggests that carrying out a regression without adjusting for group-level heterogeneity can lead to misleading results.

1.4 Group-level varying effects with dummy variables

Then, how can we estimate the linear relationship between variables adjusting for group-level heterogeneity? As we discussed in Topic 5, the simplest and most straightforward approach is to include the group-level variable as another predictor along with other predictors in the multiple

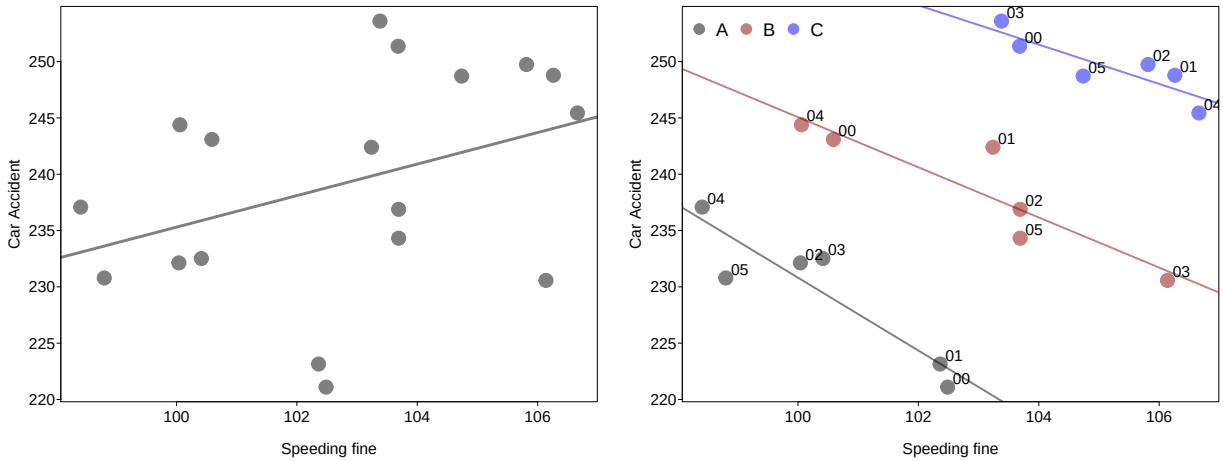


Figure 2: Car accident vs. speeding ticket, pooled & country-level

regression setting. Since the multiple regression allows us to estimate the coefficient of one variable holding other variables constant, we can estimate the impact of the main predictors on the outcome variable while accounting for the group-level difference.

One way to include a group-level variable in the regression is to use an *indicator variable* with 0, 1, which is often called a *dummy variable*. Each observation has a corresponding dummy variable indicating whether this observation is included in each group. For example, if the first observation is from Country A, the dummy variable for Country A is 1 for this observation, while the dummy variables for Country B and C are 0. Table 3 shows how to use dummy variables to indicate group-membership. This table is the same as Table 2 but with separate country and year dummy variables. The first observation, for example, is from Country A and Year 2000.

Table 3: Table 2 with country and year dummy variables.

Obs.	Accident	Fine ($)	Ctry_A	Ctry_B	Ctry_C	Y_2000	Y_2001	Y_2002	Y_2003	Y_2004	Y_2005
1	221	102	1	0	0	1	0	0	0	0	0
2	223	102	1	0	0	0	1	0	0	0	0
3	232	100	1	0	0	0	0	1	0	0	0
4	233	100	1	0	0	0	0	0	1	0	0
5	237	98	1	0	0	0	0	0	0	1	0
6	231	99	1	0	0	0	0	0	0	0	1
7	243	101	0	1	0	1	0	0	0	0	0
8	242	103	0	1	0	0	1	0	0	0	0
9	237	104	0	1	0	0	0	1	0	0	0
10	231	106	0	1	0	0	0	0	1	0	0
11	244	100	0	1	0	0	0	0	0	1	0
12	234	104	0	1	0	0	0	0	0	0	1
13	251	104	0	0	1	1	0	0	0	0	0
14	249	106	0	0	1	0	1	0	0	0	0
15	250	106	0	0	1	0	0	1	0	0	0
16	254	103	0	0	1	0	0	0	1	0	0
17	245	107	0	0	1	0	0	0	0	1	0
18	249	105	0	0	1	0	0	0	0	0	1

Since we have two different group variables in our data, we can include each of the two variables or both of them. To see how the estimated impact of speeding fine on car accidents changes in

different models, let's look into the following four regression models:

$$
\begin{aligned}
y_{ct} &= \beta_0 + \beta_1 x_{ct} + u_{ct} && (1)\\
y_{ct} &= \beta_0 + \beta_1 x_{ct} + \alpha_2 Ctry_B + \alpha_3 Ctry_C + u_{ct} && (2)\\
y_{ct} &= \beta_0 + \beta_1 x_{ct} + \gamma_2 Y_{2001} + \gamma_3 Y_{2002} + \cdots + \gamma_6 Y_{2005} + u_{ct} && (3)\\
y_{ct} &= \beta_0 + \beta_1 x_{ct} + \alpha_2 Ctry_B + \alpha_3 Ctry_C + \gamma_2 Y_{2001} + \gamma_3 Y_{2002} + \cdots + \gamma_6 Y_{2005} + u_{ct} && (4)
\end{aligned}
$$

where

y_{ct} is the car accident of country c in year t

x_{ct} is the speeding fine of country c in year t

u_{ct} is the idiosyncratic error for the observation country c in year t

$Ctry_c$ is a dummy variable (or indicator variable) for Country c: if the observation is in Country c, $C_c = 1$. Otherwise, $C_c = 0$

Y_t is a dummy variable for Year t:if the observation is in Year t, $Y_t = 1$. Otherwise, $Y_t = 0$

Eq 1 is a simple regression model for pooled data where we ignore the group-level variables. Eq 2 is a multiple regression model with the country dummy variable, while Eq 3 is a model with the year dummy variable. Finally, Eq 4 is a regression model with both the country and year dummy variables.

Note that we do not have a dummy variable for Country A and Year 2000 to avoid perfect multicollinearity (the dummy variable trap), meaning that the reference point when all dummy indicators are zero is Country A in Year 2000. Therefore, β_0 estimates the baseline car accidents for Country A and Year 2000 (a hypothetical number of car accidents when the speeding fine is zero).

This model is often called a *fixed effect* model in some disciplines.[2] Since the dummy variable takes either zero or one, these models can be written more cleanly as follows, which is a typical fixed effect model specification:

$$
\begin{aligned}
y_{ct} &= \beta_0 + \beta_1 x_{ct} + u_{ct} && (5)\\
y_{ct} &= \beta_1 x_{ct} + \alpha_c + u_{ct} \quad c = 1, 2, 3 && (6)\\
y_{ct} &= \beta_1 x_{ct} + \gamma_t + u_{ct}, \quad t = 1, \cdots, 6 && (7)\\
y_{ct} &= \beta_1 x_{ct} + \alpha_c + \gamma_t + u_{ct} && (8)
\end{aligned}
$$

α_c estimates the overall change in the number of car accidents in Country c relative to Country A, adjusting for the year-specific heterogeneity and the speeding fine. In the standard panel data analysis, this coefficient is called a country *fixed* effect because the effect is fixed across all years. Similarly, γ_t estimates the overall change in the number of car accidents in Year t relative to Year 2000, adjusting for the country-specific heterogeneity and the speeding fine. This is a year fixed effect because the effect is fixed across all countries. Finally, β_1 estimates the impact of the speeding fine on the number of car accidents, adjusting for both country- and year-fixed effects.

[2] More precisely, estimators obtained from the dummy variable regression (Least Square Dummy Variable, LSDV) are algebraically equivalent to the fixed effect estimators, which can be obtained via mean-differencing (within estimator).

Before we move to the estimation result of these models, let us examine some definitional issues involving the "fixed effect." In the above-mentioned models, fixed effect coefficients refer to coefficients that are fixed across other variables: the country fixed effects are fixed across different years while the year fixed effects are fixed across different countries. Confusingly, these fixed effect coefficients do vary by their own group variable: the country fixed effects vary by Country A, B and C, while the year fixed effects vary by different years. As we shall see, the fixed effect has a completely different definition in a Bayesian multilevel model where it refers to coefficients that do not vary by groups. To avoid unnecessary confusion involving definitional issues across different disciplines, I will try to minimize the use of terms "fixed effect" (and "random effect") and focus more on the model property itself.

Table 4: Estimation results of the four models in Eq 1 - 4.

	Outcome variable:			
	Car accidents			
	Eq 1	Eq 2	Eq 3	Eq 4
Car accident	**1.400**	**-2.438**	**1.644**	**-2.777**
	(0.864)	(0.347)	(1.052)	(0.313)
Country B		15.201		16.044
		(1.652)		(1.374)
Country C		31.537		33.123
		(2.148)		(1.850)
Year 2001			-3.202	4.320
			(8.975)	(1.689)
Year 2002			-0.461	3.650
			(8.849)	(1.630)
Year 2003			-1.361	3.319
			(8.865)	(1.637)
Year 2004			4.678	2.277
			(8.814)	(1.612)
Year 2005			-0.836	-0.138
			(8.797)	(1.604)
Constant	95.280	474.263	70.443	506.119
	(88.824)	(34.857)	(107.709)	(31.280)
Observations	18	18	18	18
R2	0.141	0.948	0.202	0.978

Table 4 summarizes the OLS regression result for all four models in Eq 1 - 4. The result for the pool regression model (Eq 1) shows that the coefficient of the speeding fine, β_1, is positive. When the country dummy variable is included (Eq 2), however, β_1 changes signs. This confirms our visual inspection of the pattern in Figure 2. In contrast, the model with the year dummy variable (Eq 3) results in a positive β_1 with a very similar magnitude of the coefficient in the pooled model Eq 1, implying that the year-level heterogeneity does not affect the estimation of the impact of speeding fine on car accidents. A similar result can be seen in the model with both year and country dummy variables (Eq 4) where the estimated coefficient β_1 is almost identical to that of the model with the country dummy variable (Eq 2). This lack of the impact of the year-level variation can be seen in Figure 3 where the relationship between car accidents and speeding fine is plotted conditional on the year group. Unlike the country group conditioning, there is still a clear positive relationship between the two variables.

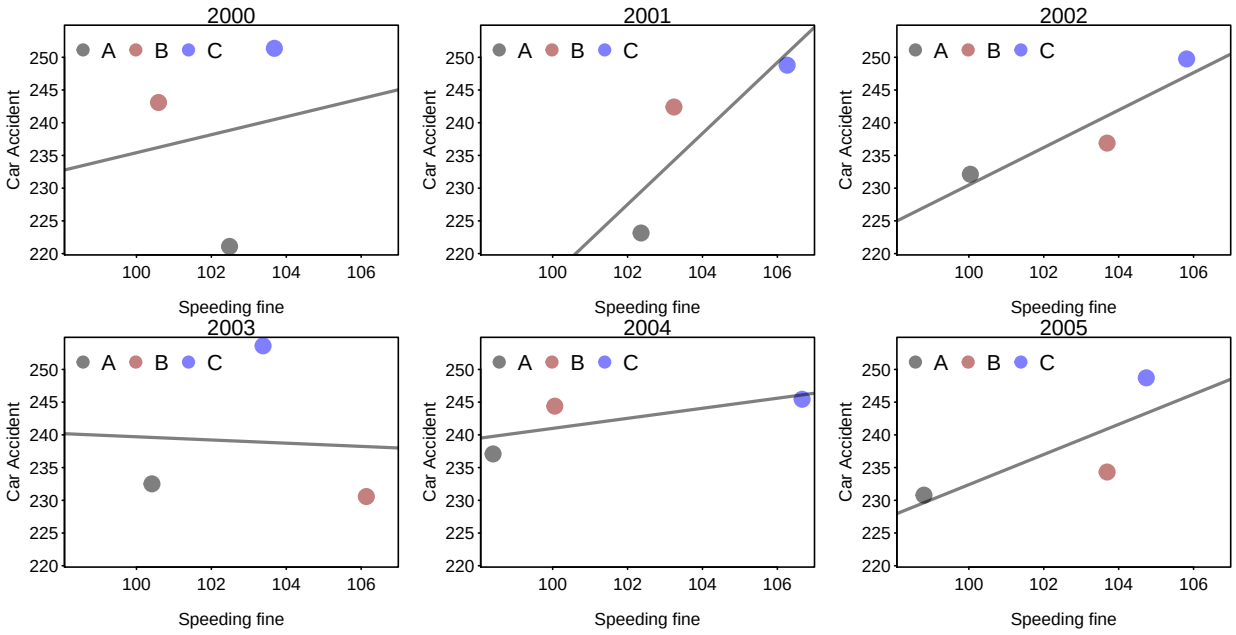


Figure 3: Car accident vs. speeding ticket, year-level

As this exercise suggests, the impact of adding group-level variables is not monolithic: it can sometimes change the sign of the main coefficient but it also can have no impact. A good understanding of the underlying data-generating process aided by a good theory can help researchers to select potentially influential group-level variables in data. However, identifying which group-level variable plays an important role is an empirical task in the end.

It is important to note that adding a group-level dummy variable allows the "intercept" to vary by group. For example, the estimated coefficients for the country dummy variable (Eq 2) are 15.201 and 31.537 for Country B and Country C, respectively. This means that the intercepts of the linear line for Country B and C move up by 15.201 and 31.537 from the baseline constant (the intercept for Country A) of 474.263. For this reason, the least square regression with a dummy variable can be understood as a type of *varying intercept models*, a general set of models where intercepts are allowed to vary by predetermined groups. As we will see in the following section, the dummy variable approach is a very special case of the varying intercept model where the variance between groups is assumed to be infinite.

2 Varying coefficient model: concepts

2.1 Some notations

A standard notation for the OLS model with a dummy variable (Eq 1-4 becomes unwieldy when we have multiple group variables to include because each observation of the variable is accompanied by all group indicators in the data. From now, we will use separate group indicators and express all variables at their individual observational level i.[3] For a varying intercept model, we use the

[3]We can also drop i and treat each variable as a vector. This is what we have assumed throughout Topics 1-6. However, we will keep i subscript in Topic 7 to give a more clear connection between the unit-level observation i and the group-level observation, c, t.

following notation:

$$y_i \quad = \quad \alpha_{t[i]} + \gamma_{c[i]} + \beta_1 x_i + \varepsilon_i \quad \text{for } i = 1, \ldots, 18, \tag{9}$$

where subscript i represents the i-th observation/unit in the data, $t[i]$ is the year group t containing unit i, and $c[i]$ is the year group c containing unit i. Same as before, α_t and γ_c are the varying intercept for the year group and the country group respectively.

2.2 Pooling, no-pooling, and partial pooling

So far, we have discussed two different approaches to group-level heterogeneity: a *complete pooling model* and *a no-pooling model*. Complete pooling assumes that there is zero variance between the subgroups and there is essentially no difference between them. Put it differently, the complete pooling assumes that the observations in each group come from exactly the same data-generating process. This assumption leads to a pooled regression in Eq 1 or Eq 5. In contrast, no pooling assumes that each group is independent of one another (infinite variance between the individual subgroups), and the observations in each group are a result of completely different data-generating processes. This assumption leads to the dummy variable regression in Eqs 2-4 or Eqs 6-8.

The varying coefficient model that we introduce in this section is a compromise of these two extremes and is often called the *partial pooling model*. In this model, the group-level variation is accounted for, but not to the extent that each group is its own isolated unit whose generative process has nothing in common with the other groups.

There are two key advantages of using partial pooling models. First, the similarity (or the difference) of observations across different groups is directly estimated from the data without assuming it beforehand. This becomes important when the group-level variability is a centerpiece of research designs. Second, the partial pooling effectively pools outlying estimates to the grand mean coefficient. This desirable property helps to avoid overfitting and leads to a lower *total* mean squared error compared to estimating each coefficient separately.

2.3 James-Stein estimator: baseball example

Then, why and how does the partial pooling minimize the mean squared error of estimation? To see this point, let us go through a famous example by Efron & Morris (1977) where authors discuss James's paradox using a baseball example. Suppose we are interested in a baseball player's true batting ability using the data on the player's observed average of success, e.g. 7 hits in 20 official times. The true batting ability is then used to project how well the player's batting record will be in the future, e.g. in the next 100 times at-bat. A common-sense approach to this seemingly simple problem is to use the observed batting average 0.350 (7/20) and predict 35 hits in the next 100 times at bat(100×0.350). However, this simple extrapolation from the observed average turns out to work rather poorly when we have three or more baseball players. That is, when we want to predict future batting average for each of the multiple players, simple extrapolation from the observed average of each player is less accurate than the alternative estimator, called the *James-Stein estimator*.

Efron & Morris (1977) dissect this paradox by examining the batting average of 18 major-league players. The first 45 times at bat in the 1970 season are used to calculate the observed batting average for each of the players, which we denote by y, a vector of 18 batting averages. A simple

extrapolation for the future batting average for players is based on this vector y. In contrast, the James-Stein estimator uses the *grand average* of the observed quantity, $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$. The key mechanism is that when a player's observed batting average is different from the grand average, it is adjusted to some degree. For example, if a player's batting average is greater than the overall mean, it needs to be reduced. We will denote this adjusted batting ability by z, a vector of 18 adjusted batting averages, which can be found through the following equation:

$$z_i = \bar{y} + c(y_i - \bar{y}) \tag{10}$$

In other words, the James-Stein estimator for the true batting ability, z, is calculated by adjusting the difference between the observed average of each player y and the grand average $\bar{y}$ by the constant factor c, which is called a shrinking factor.[4]

In the authors' calculation, the grand average $\bar{y}$ and the shrinkage factor c are 0.265 and 0.212, respectively. Therefore, if the player's observed average is .400, the James-Stein estimator for this player's true batting ability is .265 + .212(.400 - .265) =.294, which is significantly smaller than .400. The authors go on and show that the James-Stein estimator z is a better indicator of the players' true batting ability because it better predicts the future batting average for each of the players. Denoting the future batting average by θ, the prediction error is calculated via a total squared error (the quadratic loss function): $(\theta - y)^2$ and $(\theta - z)^2$, respectively. It turns out that $(\theta - y)^2 = 0.077$ and $(\theta - z)^2 = 0.022$, meaning that the James-Stein estimator has almost 4 times smaller prediction error.

This is not a fluke. James-Stein theorem proves that the estimator with the shrinkage factor, z, always performs better than the simple average y when we have 3 or more separate groups in the data. It has a remarkable implication for regression with multigroup data. The least-square regression with dummy variable we discussed above is essentially the varying intercept model where the intercept varies by the pre-determined group. The problem is that the intercept of each group is estimated separately from one another with no shrinkage factor, which, via the James-Stein theorem, leads to a higher total squared error than the estimator with a shrinkage factor.

Note that the simple observed average y is an *unbiased estimator* for the future average θ, while the estimator with a shrinkage factor, z, is a *biased estimator* by definition (See Topic 2). Here, we encounter another example of the bias-variance tradeoff where the simultaneous reduction of bias (in-sample prediction error) and variance (the mean-squared error, or the out-of-sample prediction error) is extremely difficult. However, if introducing a bit of bias can lead to a significant gain in out-of-sample prediction, having a biased estimator cannot be a concern. And, this is exactly what the James-Stein estimator does.

In the following section, we will discuss how we can rephrase the James-Stein estimator as a hierarchical model within a Bayesian framework. By relying on a Bayesian approach, we will be able to see a very clear interpretation of the shrinkage factor.

[4]Calculating the shrinking factor is important but requires a more detailed discussion. Since we will introduce another approach to shrinking factor (Bayesian hyper prior) in the following section, we will skip the calculation of James-Stein's shrinking factor but focus more on the property of the estimator itself.

3 Bayesian varying coefficient model

3.1 Varying intercept model

Model setup

This section discusses a particular specification of a varying intercept model as was introduced in Eq 9. We will focus on high-level conceptual aspects of the model than on its implementation.[5]

In line with Eqs 1-4, we will compare four different models: a base model with no varying component, and three different models with varying intercepts for a country group, for a year group, and for both country and year groups. Using a normal error model (a normal likelihood), a varying intercept model is described in the following equations:

$$
\begin{aligned}
y_i &\sim \text{Normal}\,(\mu_i, \sigma) && (11)\\
\mu_i &= \beta_0 + \beta_1 x_i && (12)\\
\mu_i &= \beta_0 + \gamma_{c[i]} + \beta_1 x_i && (13)\\
\mu_i &= \beta_0 + \alpha_{t[i]} + \beta_1 x_i && (14)\\
\mu_i &= \beta_0 + \gamma_{c[i]} + \alpha_{t[i]} + \beta_1 x_i && (15)
\end{aligned}
$$

The basic notation was discussed above in Eq 9. Two additional parameters are β_0 and σ. β_0 is the estimation of the overall intercept or the grand mean. It is a reference point for the group-varying effects, $\gamma_{c[i]}$ and $\alpha_{t[i]}$, which indicate the deviation of a specific country- and year-group intercept, respectively, from the overall intercept β_0. σ is the scale parameter of the normal distribution and can be interpreted as the overall residual of the model.

Note that there are two different types of parameters. One is the population-level parameter such as β_0 and β_1. These parameters are defined at the population-level and therefore are fixed across their own groups. The other set of parameters is the group-level parameter such as γ_c and α_t. These parameters are defined at the group-level and therefore are varying across their own groups.

Now that we set up a basic varying intercept model, let us discuss how to incorporate the "shrinkage factor" of the James-Stein estimator in our varying intercept model. Remember that the James-Stein estimator pulls group-level observations to a grand mean whose degree of pulling is determined by the shrinkage factor. In a Bayesian multilevel model, we achieve this by allowing the group-level parameters (γ_c, α_t) to be drawn from a common overall prior distribution that is governed by *hyperparameters*. Since the group-level parameters are generated from an overall distribution that is "one level higher", this model is also called a *hierarchical model*. To see this point more clearly, let us formulate a model where the group-level parameters are assumed to come from the same normal distribution with unknown scale parameter. This scale parameter is a hyperparameter that determines the degree of heterogeneity across groups and is equivalent to the shrinkage factor in the James-Stein estimator. This way, all group-level parameters are from the

[5] A detailed discussion on prior specifications and estimation strategy is not covered in this section. Also, an important discussion on the correlation between predictors and group-effects (Bafumi-Gelman estimator, Bafumi & Gelman (2006)) is not discussed in this section. These topics will be discussed in a more advanced course.

common generative process but their variation can also be accounted for. The following equations show the prior distributions for γ_c and α_t:

$$\gamma_c \sim \text{Normal}(0,\ \sigma_\gamma) \quad \text{for } c = A, B, C \tag{16}$$

$$\alpha_t \sim \text{Normal}(0,\ \sigma_\alpha) \quad \text{for } t = 2000, ..., 2005 \tag{17}$$

Eq 16 shows that the country-level varying intercepts γ_c are drawn from a normal distribution with unknown standard deviation[6], σ_γ. The same goes for α_t. As noted above, one of the key advantages of using this type of varying coefficient model is that the variance of observations across different groups, σ_γ and σ_α, is directly estimated from the data.

To be fully Bayesian, we need prior distributions on unknown parameters, $\sigma_\gamma, \sigma_\alpha$ (See Topic 3). When priors are applied to hyperparameters, they are called *hyperpriors*. The choice of the functional form of the priors requires more detailed discussions that are beyond the scope of this course. Interested students can refer to Gelman et al. (2017) for an overview on Bayesian priors. As we briefly discussed in Topic 3, many of the prior specifications are mostly motivated by computational efficiency. Also, parameter values of the prior distributions are often chosen to make the prior only weakly informative and thus make the estimation results less sensitive to prior specifications. For our exercise, we will be using the following prior specifications:

$$\beta_0 \sim \text{Student-t}\,(0, 3, 10) \tag{18}$$

$$\beta_1 \sim \text{Normal}(0,\ 10) \tag{19}$$

$$\sigma_\gamma, \sigma_\alpha, \sigma \sim \text{HalfCauchy}\,(0, 1) \tag{20}$$

The overall intercept (grand mean), α_0 is given a weakly informative prior of the Student-t distribution centered at 0, with 3 degrees of freedom, and 10 standard deviations. The hyperprior on our hyperparameters σ_γ and σ_α are given a half-Cauchy prior centered at 0 with the scale parameter 1. The model residual parameter σ has the same half-Cauchy prior.

Estimation results
The estimation results for varying intercept models are displayed in Table 5.[7] The format of the summary table follows Table 4 for easy comparison between a dummy variable approach (no pooling) and a Bayesian varying coefficient model (partial pooling).

The results for varying intercept models have three main components. The first one is the coefficients of the population-level predictors, β_0 and β_1. Both dummy variable model and Bayesian varying intercepts model (in all four models) seem to have similar estimated mean values for β_0 and β_1. Same as before, including the year-varying effect reverses the sign of β_1.

The second component is the set of coefficients for country-varying intercepts shown in Eq 13 and Eq 15. For example, γ_A in the model with the country-varying intercept Eq 13 is estimated to be -15.11 with the standard deviation of 7.634. This means that the intercept of Country A is 15.11 smaller than the overall intercept, which is estimated to be 482.24. The same logic goes for γ_B and γ_C. σ_γ is the scale parameter of the prior on γ_c, reflecting the degree of

[6]When the variable is mean-centered, we can set the mean deviation from the grand to be zero.

[7]For the posterior simulation, we use the Bayesian programming language Stan that operationalizes the Hamiltonian Monte Carlo (HMC) algorithm to efficiently compute posterior distributions of specified parameters. (Stan Development Team 2020)

Table 5: Estimation results for the four models in Eqs 12 - 15. The estimates are the mean value of the posterior distribution of each parameter. The standard deviation is in the parenthesis.

		Outcome variable:			
		Car accidents			
		Eq 12	Eq 13	Eq 14	Eq 15
Population-Level Predictors	Grand Intercept, β_0	93.86	482.24	88.07	502.08
		(94.065)	(39.683)	(96.336)	(39.387)
	Speeding fine, β_1	1.42	-2.37	1.47	-2.56
		(0.954)	(0.378)	(0.938)	(0.375)
Country-Level Predictors	Country A, γ_A		-15.11		-16.02
			(7.634)		(7.406)
	Country B, γ_B		-0.2		-0.59
			(7.578)		(7.318)
	Country C, γ_C		15.87		15.92
			(7.608)		(7.356)
	Country SD, σ_γ		16.4		16.73
			(7.32)		(7.735)
Year-Level Predictors	Year 2000, α_{2000}			-0.01	-1.13
				(2.813)	(1.383)
	Year 2001, α_{2001}			-0.59	1.02
				(2.881)	(1.407)
	Year 2002, α_{2002}			0	0.73
				(2.76)	(1.305)
	Year 2003, α_{2003}			-0.24	0.59
				(2.762)	(1.288)
	Year 2004, α_{2004}			0.99	0.13
				(2.981)	(1.284)
	Year 2005, α_{2005}			-0.14	-1.23
				(2.798)	(1.419)
	Year SD, σ_α			2.85	3.00
				(2.489)	(2.72)
	Observations	18	18	18	18
	Bayesian R2	0.16	0.93	0.2	0.95
		(0.122)	(0.019)	(0.116)	(0.019)

difference/similarity between country-varying intercepts. In the model with the country-varying intercept, it is estimated to be 16.4, which is significantly higher than the one in the model with the year-varying intercept.

The third component is the set of coefficients for year-varying intercepts shown in Eq 14 and Eq 15. As noted above, σ_α is estimated to be very small, suggesting that there is not much variation among year-varying intercepts.

The Bayesian version of R^2 is presented for each of the models. Same as the dummy variable model, it shows that including a country-varying intercept increases the model fit significantly. The fit of each model can be visually checked in Figure 4 where predicted values of car accidents, y_{rep} from the model are compared to the actual observation, y. This type of exercise is commonly known as the *posterior predictive check*. It is clear that predicted values in the model with the varying country intercept more closely match the actual observations, confirming higher Bayesian

R^2 for these models.

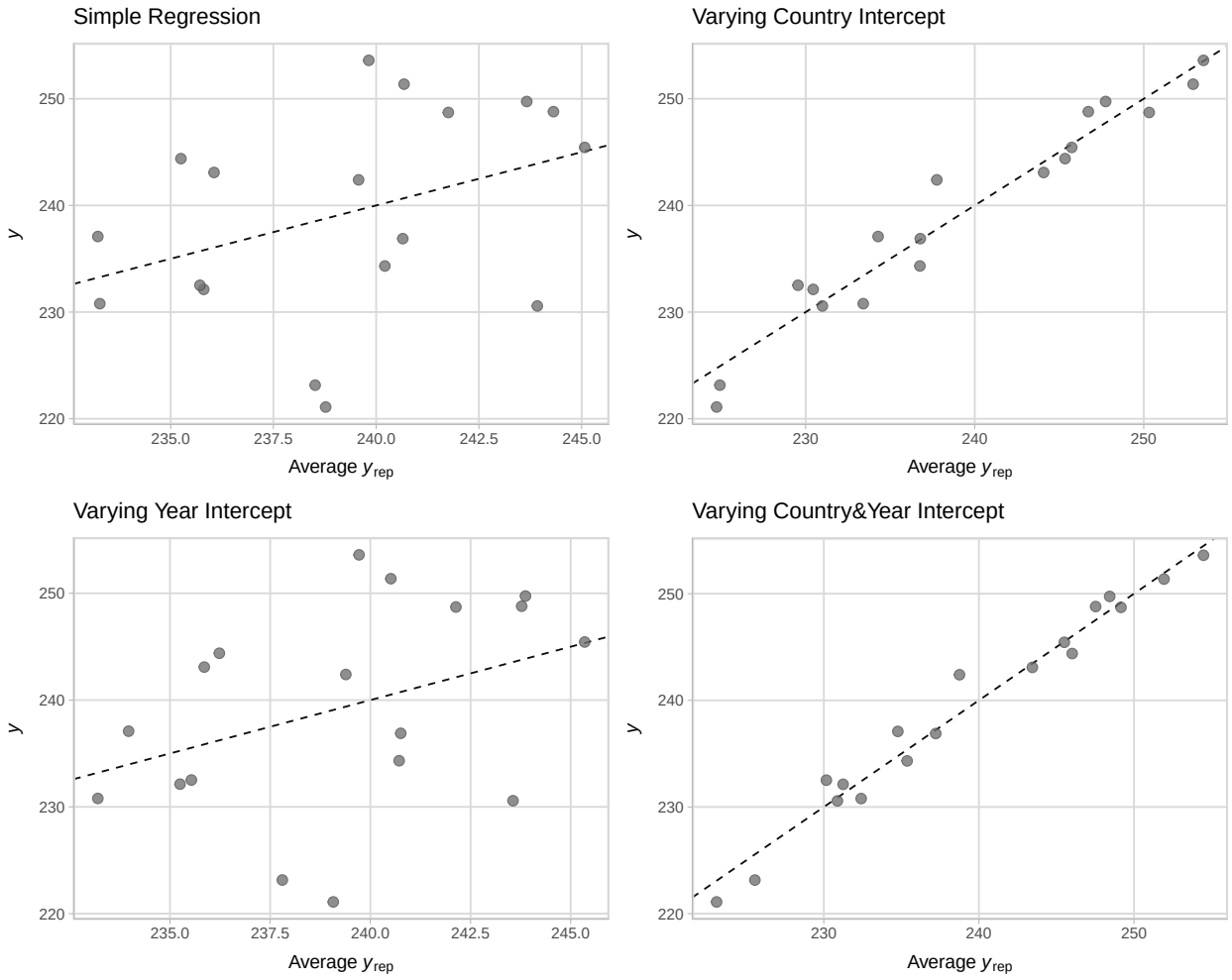


Figure 4: Predictive checking for model estimation of Eqs 12 - 15

There is one last important point we need to examine. We introduced the Bayesian multilevel model in the context of the shrinkage factor in the James-Stein estimator and showed that the shrinkage is achieved through partial pooling. Then, how can we examine to which degree shrinkage took place in our estimation? The most intuitive way to check shrinkage is to directly compare the varying coefficients from the no pooling (dummy variable model) and partial pooling models (Bayesian multilevel model). Figure 5 shows the comparison of estimated group-level intercepts for two models. The left-hand side compares the country level intercept in Eq 1 and Eq 12 while the right-hand side compares the year level intercept in Eq 3 and Eq 14. Intercepts in no pooling and partial pooling models are in black and in red, respectively.

It is clear that the partial pooling models estimate the group-level intercepts closer to one another, meaning that country- and year-varying intercepts are pulled toward the grand mean. Note that when the estimates of the group-level coefficients are nosier, the degree of shrinkage is higher so that they are pulled more strongly toward the grand mean. This can be seen in the year

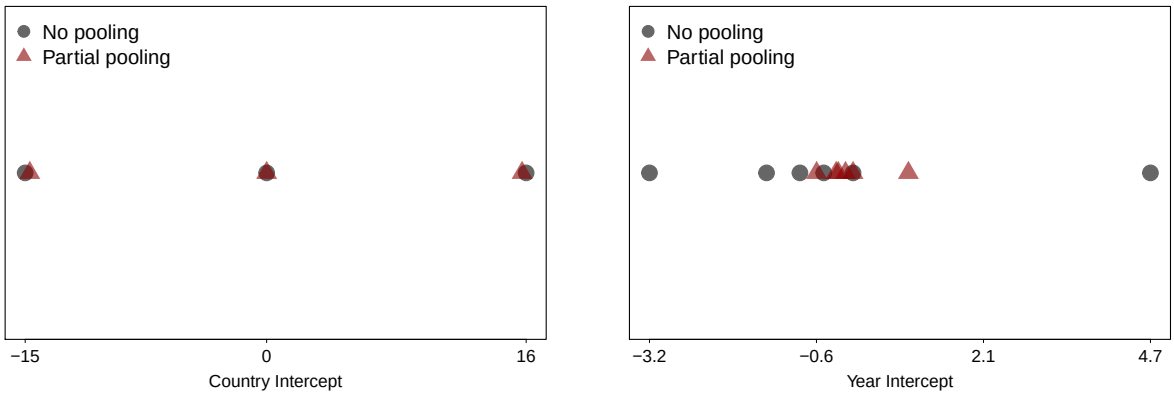


Figure 5: No pooling vs. partial pooling

varying intercept where the estimation error (or the standard deviation of the posterior distribution) tends to be higher.

3.2 Varying slope model

In the linear regression setting, we can also allow the slope coefficients to vary by groups. To showcase effectiveness of a varying slope model, we will use a different data set because the car accident data does not have much variation in the group-level slope as shown in Figures 2. The new (simulated) data is shown in Figure 6. The left-hand side plot shows a simple linear relationship between x and y while the right-hand side plot shows the same relationship conditional on three different groups. It is clear that each group has different slopes and intercepts.

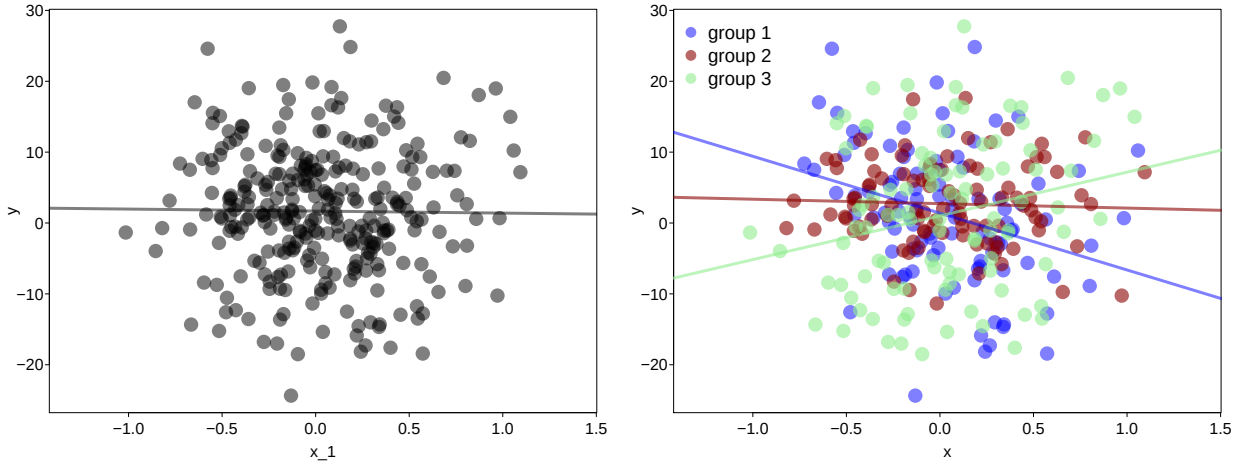


Figure 6: x vs. y, pooled & group-level

To understand this data, we will use a varying coefficient model where both intercepts and slopes vary by group. The model can be written as follows:

$$y_i \sim \text{Normal}(\mu_i, \sigma) \quad (21)$$
$$\mu_i = \alpha_0 + \alpha_{j[i]} + (\beta_1 + \beta_{j[i]})x_i \quad (22)$$

where α_0 and β_1 are the population-level intercept and slope while α_j and β_j are the varying intercept and slope for group j. As before, group-varying effects are estimated as the deviation from the population effect.

The prior specification is the following:

$$\alpha_0 \sim \text{Normal}(0,\ 10) \quad (23)$$
$$\beta_1 \sim \text{Normal}(0,\ 5) \quad (24)$$
$$\sigma \sim \text{HalfCauchy}(0, 1) \quad (25)$$
$$\alpha_j \sim \text{Normal}(0,\ \sigma_\alpha) \quad (26)$$
$$\beta_j \sim \text{Normal}(0,\ \sigma_\beta) \quad (27)$$
$$\sigma_\alpha, \sigma_\beta \sim \text{HalfCauchy}(0, 1) \quad (28)$$

where Eq 26 and Eq 27 represent the shrinkage prior on the varying intercept and slope. Note that a more sophisticated model can include a covariance structure between α_j and β_j since these parameters move within the same group. Table 6 summarizes the estimation results. Since we are not comparing different models as we did in Table 5, we provide more detailed information about the posterior distribution of each parameter in the model, including the 2.5% and 97.5% of the posterior distribution.

Table 6: Estimation result for the model in Eq.22

	Parameter	Mean	SD	2.5% Q	97.5% Q
Population-Level Predictors	α_0	1.642	1.59	-2.019	4.787
	β_1	-0.698	6.19	-13.543	11.732
Varying Intercept	α_{group1}	-0.07	1.59	-3.404	3.57
	α_{group2}	0.604	1.679	-2.229	4.881
	α_{group3}	-0.282	1.598	-3.799	3.172
	σ_α	1.908	2.363	0.047	8.508
Varying Slope	β_{group1}	-6.397	6.382	-19.795	6.36
	β_{group2}	0.155	6.239	-13.063	12.944
	β_{group3}	6.029	6.311	-6.562	19.476
	σ_β	9.204	5.362	3.106	22.311
Residual Error	σ	8.701	0.365	8.012	9.442

As predicted from Figure 6, the group-level slope is estimated to significantly vary by each group. For example, the slope of the linear relationship between x and y in Group 1 is -6.397 while that in Group 3 is 6.029.

Finally, let us discuss the effect of partial pooling in the model. As we discussed in Topic 5, the varying slope can also be formulated by adding an interaction of the group level variable and the

main predictor, which in essence adds dummy variables in the slope coefficient, $(\beta_1 + \beta_j)x$. This is exactly the same expression as in Eq. 22 but without any common generative process of β_j. That is, linear regression with interactions (with the group-level factor) is the no-pooling counterpart to our Bayesian partial-pooling model.[8] Figure 7 shows a comparison of the group varying slopes in no-pooling and partial pooling. As predicted, there is some degree of shrinkage in the partial pooling model.

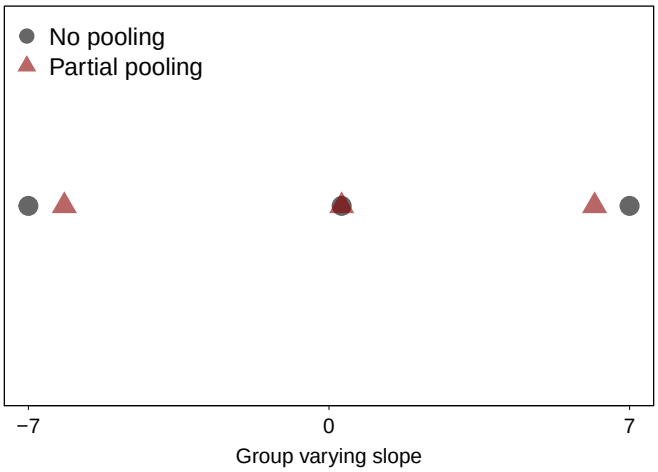


Figure 7: No pooling vs. partial pooling

[8]Equivalently, we can use a very wide distribution whose variance is extremely large and use this as a prior distribution on the group-varying coefficients.

References

Bafumi, J. & Gelman, A. E. (2006), 'Fitting multilevel models when predictors and group effects correlate'.

Efron, B. & Morris, C. (1977), 'Stein's paradox in statistics', *Scientific American* **236**(5), 119–127.

Gelman, A., Simpson, D. & Betancourt, M. (2017), 'The prior can often only be understood in the context of the likelihood', *Entropy* **19**(10), 555.

Stan Development Team (2020), 'RStan: the R interface to Stan'. R package version 2.21.2. **URL:** *http://mc-stan.org/*