PROBABILITY AND STATISTICS A

Teruo Nakatsuma

Spring Semester 2021

Faculty of Economics, Keio University

Aims Of This Course

1. Learn basic principles of Bayesian learning.
2. Learn how to conduct statistical inference (point estimation, interval estimation, model selection) in the Bayesian way.
3. Learn basic principles of Markov chain Monte Carlo (MCMC) methods.
4. Hands-on practice of Python and PyMC.

Reading List i

1. Introduction to Bayesian statistics
    - Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B. (2013). *Bayesian Data Analysis*, 3rd ed., Chapman & Hall/CRC.
    - Greenberg, E. (2013). *Introduction to Bayesian Econometrics*, 2nd ed., Cambridge University Press.
2. Advanced topics in Bayesian statistics
    - Chan, J., Koop, G., Poirier, D.J. and Tobias, J.L. (2019). *Bayesian Econometric Methods*, 2nd ed., Cambridge University Press.

Reading List ii

- Durbin, J. and Koopman, S.J. (2012). *Time Series Analysis by State Space Methods*, 2nd ed., Oxford University Press.
- Prado, R. and West, M. (2010). *Time Series: Modeling, Computation, and Inference*, Chapman & Hall/CRC.
- Rossi, P.E., Allenby, G.E. and McCulloch, R. (2005). *Bayesian Statistics and Marketing*, Wiley.

3. PyMC

   - Davidson-Pilon, C. (2016). Bayesian Methods for Hackers: Probabilistic Programming and Bayesian Inference, Addison-Wesley.

Reading List iii

- Martin, O. (2018). Bayesian Analysis with Python, 2nd ed., Packt Publishing.

4. Markov chain Monte Carlo (MCMC)
    - Robert, C.P. and Casella, G. (2004). *Monte Carlo Statistical Methods*, 2nd ed., Springer.
5. Classics
    - Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed., Springer.
    - Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*, Wiley.

Python

- Python is a high-level programming language.
- Designed by Guido van Rossum
- Released in 1991
- Python is popular.
    - Ranking 1
    - Ranking 2

Why Python?

- It is free.
- It is slow in execution but highly manageable.
- Python codes are arguably more readable than other programming languages.
- Numerous packages (collections of functions for specific purposes) have been developed for Python.
- Most of them are free and written in faster programming languages such as C/C++.

How To Obtain Python

- The official Python website
- Unfortunately, the plain Python does not include any useful tools.
- Distributions for scientific computing
    - Anaconda
    - ActivePython
    - Enthought Deployment Manager
- Online service
    - Google Colaboratory

REPL (Read-Eval-Print-Loop)

REPL refers to a computer programming environment that allows users to write codes (Read), execute them instantly (Eval), and confirm the results on the screen (Print).

- Terminal-based REPL — **IPython**
- Browser-based REPL — **Jupyter Notebook**

IDE (Integrated Development Environment)

IDE is an application that integrates an editor, a debugger, a profiler and other useful tools for developers.

- Spyder
- PyCharm
- Visual Studio Code

Basic Packages

- NumPy — n-dimensional array and mathematical functions (`https://www.numpy.org`)
- SciPy — functions for scientific computing (`https://www.scipy.org`)
- Matplotlib — 2D/3D plotting (`https://matplotlib.org`)
- Pandas — data structure (`https://pandas.pydata.org`)

PyMC

PyMC (`https://docs.pymc.io/index.html`) is a Python package for Bayesian MCMC computation. Unlike other tools such as Stan (`https://mc-stan.org`), PyMC is specifically designed for Python and is well integrated with Python and NumPy. So you can write a very *Pythonic* code to perform MCMC computation.

Reference: Salvatier J., Wiecki, T.V. and Fonnesbeck, C. (2016). “Probabilistic Programming in Python Using PyMC3”, *PeerJ Computer Science*, 2:e55.

Population, Sample, Parameter

1. In statistics, the population is any subject (not necessarily a group) which a researcher tries to analyze.
2. The sample is a collection of data related to the population. In a typical situation, the sample is assumed to be randomly and independently extracted from the population.
3. The parameter represents a property of the population to be analyzed. The parameter is unknown to the researcher.
4. The goal of statistics is to obtain useful insights about the parameter of the population with the sample extracted from it.

Population Distribution

We regard the population as a probability distribution and call it the population distribution. Then we can interpret the sample as a set of random variables following the population distribution, and the parameters as variables which determine the “shape” of the population distribution. Let $D = (x_1, \dots, x_n)$ denote the sample, and θ denote the parameter of the population distribution. To indicate that the shape of the population distribution depends on θ, the population p.m.f. or p.d.f. is denoted by $p(x_i|\theta)$ where each x_i $(i = 1, \dots, n)$ is called an observation and supposed to be a realized value of the random variable following the population distribution. n is often called the sample size.

Example: Bernoulli Distribution

Consider a random variable X_i $(i = 1, \ldots, n)$ such that

$$X_i = \begin{cases} 1, & \text{Head is obtained;} \\ 0, & \text{Tail is obtained,} \end{cases}$$

and suppose $\mathbf{P}(X_i = 1) = \theta$ and $\mathbf{P}(X_i = 0) = 1 - \theta$. Then X_i follows the Bernoulli distribution with the p.m.f.:

$$p(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}, \quad x_i = 0, 1.$$

Likelihood

Suppose the sample $\boldsymbol{D} = (x_1, \ldots, x_n)$ are taken from a population distribution where the parameter $\boldsymbol{\theta}$ is a set of unknown parameters. The joint p.m.f. or the joint p.d.f. of D is denoted by

$$p(D|\theta) = p(x_1, \ldots, x_n|\theta).$$

In particular, if observations are independent of each other,

$$p(D|\theta) = p(x_1|\theta) \times \cdots \times p(x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta).$$

When we regard $\boldsymbol{p(D|\theta)}$ as a function of $\boldsymbol{\theta}$, it is called the likelihood or likelihood function.

Example: Bernoulli Distribution

Suppose $D = (x_1, \ldots, x_n)$ is independently generated from the same Bernoulli distribution. Then

$$p(D|\theta) = \prod_{i=1}^{n} p(x_i|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^y (1-\theta)^{n-y},$$

where $y = \sum_{i=1}^{n} x_i$.

Suppose we have $(x_1, x_2, x_3, x_4, x_5) = (1, 0, 1, 1, 1)$. The value of $p(D|\theta)$ depends on the value of θ.

θ	0.1000	0.2000	0.3000	0.4000	0.5000	0.6000	0.7000	0.8000	0.9000
$p(D\|\theta)$	0.0001	0.0013	0.0057	0.0154	0.0312	0.0518	0.0720	0.0819	0.0656

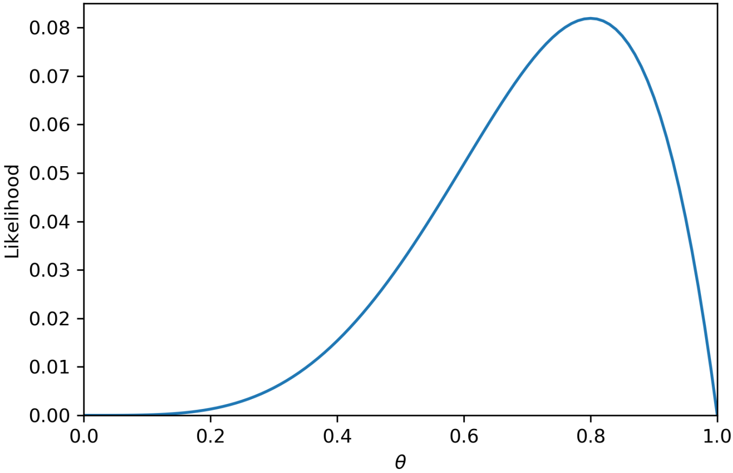

Figure 1: The likelihood of θ in the Bernoulli distribution

Interpretation Of The Likelihood

Given the sample D, the likelihood $p(D|\theta)$ is regarded as a kind of “plausibility” of a specific value of θ.

For example, the likelihood of $\theta = 0.9$ is 0.0656 while that of $\theta = 0.4$ in the previous example is 0.0154. We may say that 0.9 is about 4 times more plausible than 0.4 as the true value of θ.

To make comparison between two competing values of θ, say θ_0 and θ_1, we introduce the likelihood ratio:

$$\text{likelihood ratio} = \frac{p(D|\theta_0)}{p(D|\theta_1)}.$$

Prior Knowledge On Parameters

In practice, researchers often have information on unknown parameters before they start analysis. For example,

- $\boldsymbol{\theta}$ must take a value between 0 and 1 because it is probability;
- in case of tossing a coin, $\boldsymbol{\theta}$ is supposed to be 50% if the coin is fair.

In Bayesian statistics, we construct a distribution of unknown parameters that reflect our prior knowledge on their true values. This is call the prior distribution. Let $\boldsymbol{p(\theta)}$ denote the prior distribution.

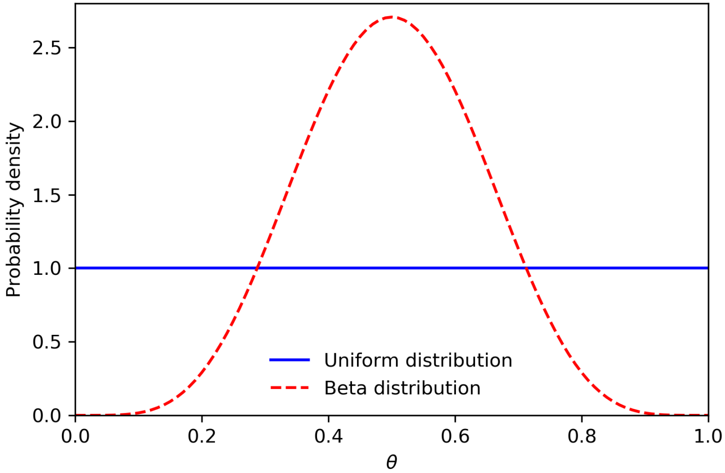

Figure 2: Prior distributions of $\boldsymbol{\theta}$ in the Bernoulli distribution

The uniform distribution Uniform(a, b) is

$$p(x|a,b) = \begin{cases} \frac{1}{b-a}, & (a \leqq x \leqq b); \\ 0, & (\text{otherwise}). \end{cases}$$

In the above figure, we set $a = 0$ and $b = 1$.

The beta distribution Beta(α, β) is

$$p(x|\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \ 0 \leqq x \leqq 1.$$

where $B(\alpha, \beta)$ is the beta function:

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx.$$

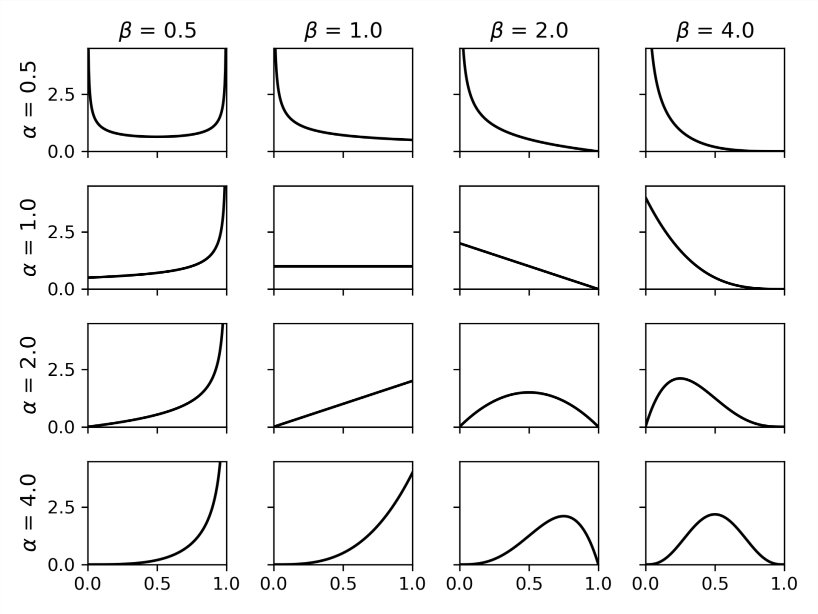

Figure 3: Beta distributions with various (α, β)

Bayes' Theorem And Posterior Distribution

Suppose $p(\theta, D)$ is the joint distribution of θ and D. Using the definition of the conditional distribution, we have

$$p(\theta, D) = p(\theta|D)p(D) = p(D|\theta)p(\theta).$$

Arranging the middle and the right-hand side of the above equation, we have

Bayes' Theorem

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}.$$

- The above formula is called Bayes' theorem.
- $p(\theta|D)$ is called the posterior distribution.
- $p(D)$ is called the normalizing constant.

Marginal Likelihood

By marginalizing the joint distribution $p(\theta, D)$, we have

$$p(D) = \int p(\theta, D) d\theta = \int p(D|\theta)p(\theta)d\theta.$$

Thus $p(D)$ is interpreted as "averaged likelihood" in terms of the prior $p(\theta)$. In this sense, $p(D)$ is called the marginal likelihood.

Then Bayes' theorem is rewritten as

Bayes' Theorem (Alternative Form)

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta} \propto p(D|\theta)p(\theta).$$

We can ignore $p(D)$ since it does not depend on θ.

Example: Bernoulli Distribution

Suppose the prior distribution is Beta(α_0, β_0).

The posterior distribution of θ is given by

$$\begin{aligned} p(\theta|D) &\propto p(D|\theta)p(\theta) \\ &\propto \theta^y(1-\theta)^{n-y} \times \theta^{\alpha_0-1}(1-\theta)^{\beta_0-1} \\ &\propto \theta^{y+\alpha_0-1}(1-\theta)^{n-y+\beta_0-1} \\ &\propto \theta^{\alpha_\star-1}(1-\theta)^{\beta_\star-1}, \\ \alpha_\star = y + \alpha_0&, \quad \beta_\star = n - y + \beta_0. \end{aligned}$$

This is the beta distribution Beta$(\alpha_\star, \beta_\star)$.

Bayesian Learning

Bayes' theorem is rearranged as

$$\frac{p(\theta|D)}{p(\theta)} = \frac{p(D|\theta)}{p(D)}.$$

Therefore

$$\frac{p(\theta|D)}{p(\theta)} \gtreqless 1 \quad \text{if and only if} \quad \frac{p(D|\theta)}{p(D)} \gtreqless 1.$$

Since

$$\begin{cases} \dfrac{p(\theta|D)}{p(\theta)} > 1, & \text{plausibility of } \theta \text{ is increased;} \\ \dfrac{p(\theta|D)}{p(\theta)} < 1, & \text{plausibility of } \theta \text{ is decreased,} \end{cases}$$

The plausibility of θ depends on whether $p(D|\theta)$ is higher than $p(D)$.

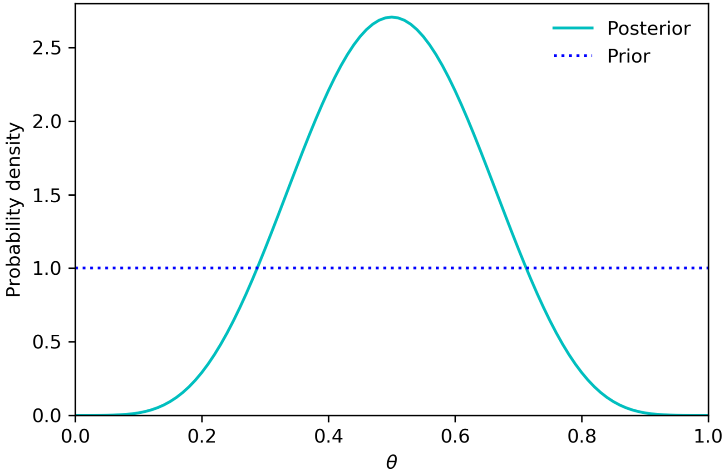

Figure 4: Posterior distributions of the probability of success

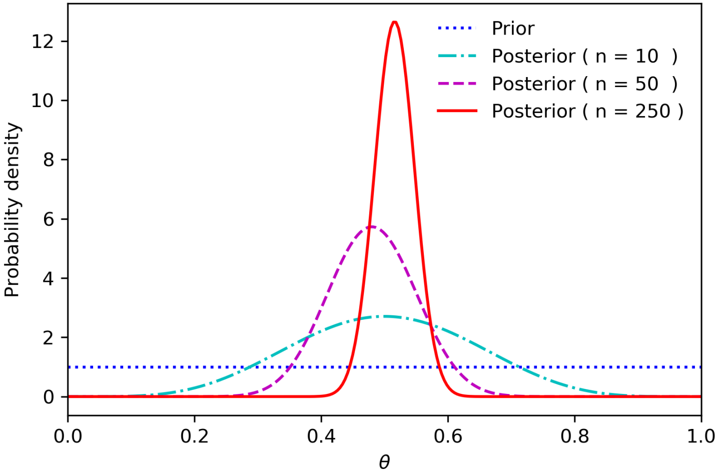

Figure 5: Sequential updating of the posterior distribution

Likelihood Principle

Likelihood Principle

Information on unknown parameters $\boldsymbol{\theta}$ brought by a sample X is entirely contained in the likelihood $\boldsymbol{p(X|\theta)}$. Furthermore, suppose that X and Y are samples dependent on the same $\boldsymbol{\theta}$ and the likelihood of X and Y are proportional to each other, i.e.,

$$p(X|\theta) = \mathcal{K} \cdot p(Y|\theta) \quad \text{for all } \theta \text{ and some } \mathcal{K} > 0.$$

Then X and Y bring the same information on $\boldsymbol{\theta}$ and the resulting inference should be the same.

This is called the likelihood principle and is regarded as the pillar of statistical inference by Bayesian statisticians.

Example: Exit Poll i

On the day of an election, news media often take an exit poll to forecast which candidate/party will win the election. Suppose there are only two candidates on the ballot in this electoral district and one of them must be elected.

One candidate belongs to the ruling party, labeled as R, and the other belongs to the opposition party, labeled as O. Pollsters ask “Who did you vote for?” to voters who just

Example: Exit Poll ii

get out of voting places in this district. They ask n voters in total and the answer by the i-th voter is recorded as

$$x_i = \begin{cases} 1, & \text{Voted for Candidate R;} \\ 0, & \text{Voted for Candidate O.} \end{cases}$$

Results of this exit poll are collected into the data set $D = (x_1, \ldots, x_n)$.

Suppose R's true vote share is θ ($0 \leqq \theta \leqq 1$).

Since pollsters randomly encounter a voter who voted for R, the probability that the i-th voter did vote for R is θ.

Example: Exit Poll iii

Furthermore $x_1, \ldots x_n$ are supposed to be independent. Therefore each outcome of the exit poll follows a Bernoulli distribution. Hence the likelihood of $\boldsymbol{\theta}$ given D is

$$\begin{aligned} p(D|\theta) &= \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} \\ &= \theta^{y} (1-\theta)^{n-y}, \quad y = \sum_{i=1}^{n} x_i. \end{aligned}$$

We can derive the posterior distribution of $\boldsymbol{\theta}$ as explained previously.

Example: Exit Poll iv

Alternatively, suppose a researcher is informed that pollsters asked n voters independently and found y voters who voted for R. In this situation, the researcher is not aware of exact composition of $D = (x_1, \ldots, x_n)$ but only knows a summary statistic y. Then y follows

$$p(y|\theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y},$$

which is also regarded as the likelihood of θ.

The above likelihood is proportional to the one in the previous slide. Thus the posterior distribution of θ is the same as before.

Example: Lindley and Phillips (1976)

Here we keep using the exit poll as an example. Consider two pollsters who used different polling methods:

1. Pollster A questioned 12 voters and 9 voters replied that they voted for Candidate R.
2. Pollster B questioned voters until she got 3 voters who voted for Candidate O. She questioned 12 voters to get the result.

Two Likelihoods, One Information

For A, the number of voters who voted for R follows a binomial distribution:

$$p(X_A|\theta) = \binom{12}{9}\theta^9(1-\theta)^3.$$

For B, on the other hand, it follows a negative binomial distribution. Thus we have

$$p(X_B|\theta) = \binom{11}{2}(1-\theta)^3\theta^9.$$

Since $p(X_A|\theta) \propto p(X_B|\theta) \propto \theta^9(1-\theta)^3$, the likelihood principle implies that the inference on θ should be the same.

In fact, the posterior distribution of θ is identical for both pollsters when we apply the same prior distribution.

Paradox On The P-Value

- The probability that the pollster finds 9 or more voters who voted for R, $\mathbf{P}(X \geqq 9)$, does depend on how she collects the data.
- Suppose $\theta = 1/2$. For A, $\mathbf{P}(X \geqq 9)$ is 7.30%. For B, $\mathbf{P}(X \geqq 9)$ is 3.27%.
- If the significance level is set at 5%, B would conclude that R has won while A would conclude the opposite. This is a classic example why the use of the P-value such as $\mathbf{P}(X \geqq 9)$ is sometimes misleading.

$\mathbf{P}(X \geqq x)$	X_A	X_B
0	1.0000	1.0000
1	0.9998	0.8750
2	0.9968	0.6875
3	0.9807	0.5000
4	0.9270	0.3438
5	0.8062	0.2266
6	0.6128	0.1445
7	0.3872	0.0898
8	0.1938	0.0547
9	0.0730	0.0327
10	0.0193	0.0193
11	0.0032	0.0112
12	0.0002	0.0065

Bayesian Inference With The Posterior Distribution

The posterior distribution $\boldsymbol{p(\theta|D)}$ embodies all available information about unknown parameters, $\boldsymbol{\theta}$. When the number of parameters to be analyzed is relatively small, displaying graphs of all (marginal) posterior distributions may be sufficient to convey useful insights on the parameters to readers.

However, when we need to analyze many parameters, it is impractical and pointless to show all graphs on the parameters in an article or report. In practice, we calculate and report several “summary statistics” that show us key characteristics of the posterior distribution. We call them the posterior statistics.

Point Estimation

On many occasions, we need to report one particular value of the parameter we regard as the most plausible guess. This type of value is called an estimate and a procedure to obtain an estimate is called point estimation.

In Bayesian statistics, an estimate of the parameter is defined as a value that minimizes the expected loss.

$$\begin{aligned} \delta_\star &= \arg\min_\delta \mathrm{E}_\theta[L(\theta, \delta)|D] \\ &= \arg\min_\delta \int_\Theta L(\theta, \delta) p(\theta|D) d\theta, \end{aligned}$$

where L is the loss function and Θ is a set of all possible values of θ (parameter space). In case of the Bernoulli probability, $\Theta = \{\theta : \ 0 \leqq \theta \leqq 1\}$.

Examples Of Loss Functions

loss function	$L(\theta, \delta)$	point estimate
quadratic loss	$(\theta - \delta)^2$	posterior mean
absolute loss	$\|\theta - \delta\|$	posterior median
0–1 loss	$1 - 1_{\theta}(\delta)$	posterior mode

where $1_{q}(\delta)$ is the indicator function such that

$$1_{\theta}(\delta) = \begin{cases} 1, & (\delta = \theta), \\ 0, & (\delta \neq \theta). \end{cases}$$

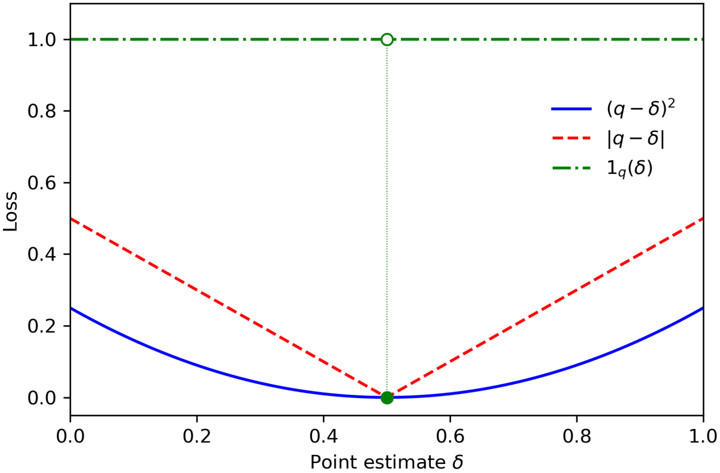

Figure 6: Examples of loss functions

Mean, Median, Mode

The mean of the distribution is the weighted average of all possible values $\boldsymbol{\theta}$ may take, i.e.,

$$\mathbf{E}_{\theta}[\theta|D] = \int_{\Theta} \theta p(\theta|D) d\theta.$$

The median of the distribution is a point that divides the distribution in half, i.e.,

$$\mathbf{P}(\theta \leqq \mathbf{Median}_{\theta}|D) = 50\%.$$

The mode of the distribution is the highest point of the density, i.e.,

$$\mathbf{Mode}_{\theta} = \arg\max_{\Theta} p(\theta|D).$$

Remarks On Point Estimation

1. A point estimate (mean, median, mode) is merely a representative point in the posterior distribution. So it is by no means the true value of the parameter in any sense.
2. The mode is not necessarily located in the center of the posterior distribution. You must be careful about using the mode as a point estimate.
3. A Bayesian point estimate is not a random variable.
4. An estimator in the frequentist approach, on the other hand, is supposed to be a random variable.
5. Some researchers use the formula of a Bayesian point estimate as a frequentist estimator. This is often called a Bayes estimator.

Posterior Probability

The probability that the true value of $\boldsymbol{\theta}$ is within a region in the parameter space, $S_0 \subset \boldsymbol{\Theta}$, is given by

$$\mathbf{P}(\theta \in S_0|D) = \int_{S_0} p(\theta|D)d\theta.$$

Such a probability is often called the posterior probability.

When the region is an interval, $S_0 = \{\theta : \ a \leqq \theta \leqq b\}$, we have

$$\mathbf{P}(a \leqq \theta \leqq b|D) = \int_a^b p(\theta|D)d\theta.$$

Credible Interval (CI)

It is tempting to state that the true value of the parameter must be within an interval with very high posterior probability (say 95%). However, there exist infinitely many intervals with 95% probability because the posterior distribution of the parameter is continuous. Thus we need extra conditions to pin down a unique interval with high posterior probability.

The credible interval of $\boldsymbol{\theta}$ is an interval $\boldsymbol{[a_c, b_c]}$ such that

1. $\mathbf{P}(a_c \leqq \theta \leqq b_c|D) = 1 - c$,
2. $\mathbf{P}(\theta < a_c|D) = \frac{c}{2}$ and $\mathbf{P}(\theta > b_c|D) = \frac{c}{2}$.

Set $c = 0.05$ for the 95% CI.

Highest Posterior Density Interval (HPDI)

The highest posterior density interval of θ is an interval $[a_c, b_c]$ such that

1. $\mathbf{P}(a_c \leqq \theta \leqq b_c | D) = 1 - c$,
2. for any pair (θ, θ') such that $\theta \in [a_c, b_c]$ and $\theta' \notin [a_c, b_c]$, $p(\theta|D) > p(\theta'|D)$ must hold.

In particular, if the distribution is unimodal (it has the unique mode), the HPDI must satisfy

$$\begin{aligned} \mathbf{P}(a_c \leqq \theta \leqq b_c | D) = 1 - c, \\ p(a_c|D) = p(b_c|D). \end{aligned}$$

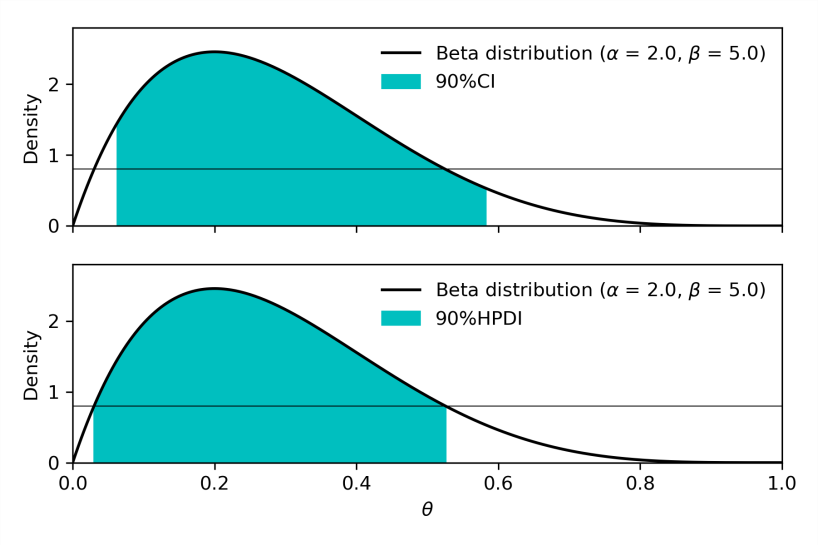

Figure 7: Comparison between CI and HPDI

Remarks On Interval Estimation

1. Both ends of CI or HPDI are not random.
2. The confidence interval in the frequentist approach, on the other hand, is a randomly shifting interval.
3. In Bayesian statistics, the posterior probability is your degree of belief. So you can say “I am 95% certain that the true value of $\boldsymbol{\theta}$ is located in the 95% HPDI.”
4. The 95% confidence interval may capture the true value of the parameter with probability 95%. This probability “95%” is interpreted as the frequency with which the confidence interval will succeed in capturing the true value.

Hypotheses On Parameters

In statistics, either Bayesian or frequentist, a hypothesis on the parameters is a region or interval where the true value of the parameter is supposed to be located. For example,

- $\{\boldsymbol{\theta} : \ 0.5 \leqq \boldsymbol{\theta} \leqq 1\}$,
- $\boldsymbol{\theta} = 0.5$,
- $\boldsymbol{\theta} \neq 0.5$
  $\Leftrightarrow \ \{\boldsymbol{\theta} : \ 0 \leqq \ \boldsymbol{\theta} \ < 0.5\} \cup \{\boldsymbol{\theta} : \ 0.5 < \boldsymbol{\theta} \ \leq \ 1\}$.

In general, a hypothesis H_i under which the true value of $\boldsymbol{\theta}$ is located in a region $S_i \subset \boldsymbol{\Theta}$ is expressed as

$$H_i : \ \boldsymbol{\theta} \in S_i, \quad i = 0, 1, 2, \ldots$$

Hypothesis Testing

In Bayesian statistics, plausibility of a hypothesis is measured by the posterior probability that the true value of $\boldsymbol{\theta}$ is located in S_i, that is,

$$\mathbf{P}(H_i|D) = \mathbf{P}(\theta \in S_i|D) = \int_{S_i} p(\theta|D)d\theta.$$

Competing hypotheses can be compared by using the posterior odds ratio:

$$\text{Posterior odds ratio} = \frac{\mathbf{P}(H_i|D)}{\mathbf{P}(H_j|D)}, \quad i \neq j.$$

Unlike the frequentist approach, Bayesian hypothesis testing does not involves the level of significance, the power and that dreaded P-value!

Bayes Factor

One catch of the posterior odds ratio is that it is affected by the prior information. If the prior information is biased in favor of one hypothesis, the posterior odds ratio is also biased for that hypothesis.

To control the impact of the prior information, the Bayes factor is often used. It is defined as

$$\text{Bayes factor} = \mathbf{B}_{ij} = \frac{\mathbf{P}(H_i|D)}{\mathbf{P}(H_j|D)} \div \frac{\mathbf{P}(H_i)}{\mathbf{P}(H_j)},$$

where $\mathbf{P}(H_i) = \int_{S_i} p(\theta)d\theta$ and $\mathbf{P}(H_i)/\mathbf{P}(H_j)$ is called the prior odds ratio. Note that the Bayes factor is equivalent to the posterior odds ratio if the prior odds ratio is one.

Scale Of Bayes Factor By Jeffreys (1961)

We compare H_i against H_j $(i \neq j)$. We suppose H_i is the hypothesis we keep unless we have no strong evidence against it. H_j, on the other hand, is the hypothesis we want to check whether it is supported by the evidence.

Rank	Bayes factor $\mathbf{B}_{ij}$	Support for H_j
0	$0 < \log_{10}(\mathbf{B}_{ij})$	Rejected
1	$-\frac{1}{2} < \log_{10}(\mathbf{B}_{ij}) < 0$	Barely worth mentioning
2	$-1 < \log_{10}(\mathbf{B}_{ij}) < -\frac{1}{2}$	Substantial
3	$-\frac{3}{2} < \log_{10}(\mathbf{B}_{ij}) < -1$	Strong
4	$-2 < \log_{10}(\mathbf{B}_{ij}) < -\frac{3}{2}$	Very strong
5	$\log_{10}(\mathbf{B}_{ij}) < -2$	Decisive

Two-Sided Test

On some occasions, we need to check whether the true value of θ is exactly equal to a particular value, say 0.5 (θ must be 0.5 if the coin is fair). For this purpose, we need to compare $H_0: \ \theta = 0.5$ against $H_1: \ \theta \neq 0.5$.

As a general setup, we consider

$$\begin{cases} H_0: & \theta = \theta_0, \\ H_1: & \theta \neq \theta_0. \end{cases}$$

For these hypothesis, however, it is meaningless to construct the Bayes factor because

$$\mathbf{P}(\theta = \theta_0) = \mathbf{P}(\theta = \theta_0 | D) = 0,$$

and prior and posterior odds ratio are identical to zero.

Spike-And-Slab Prior

To avoid this problem, we introduce a spike-and-slab prior:

$$p(\theta) = p_0\delta(\theta - \theta_0) + (1 - p_0)f(\theta), \quad 0 < p_0 < 1,$$

where $f(\cdot)$ is a continuous distribution of θ and $\delta(\cdot)$ is the Dirac delta function such that

- for any continuous function $g(x)$, $\int_{-\infty}^{\infty} g(x)\delta(x)dx = g(0)$;
- $\int_{-\infty}^{\infty} \delta(x)dx = 1$;
- $\delta(x) = 0$ only if $x \neq 0$.

Savage-Dickey Density Ratio

With the spike-and-slab prior, the prior odds ratio is $\frac{p_0}{1-p_0}$ and the posterior odds ratio is

$$\text{Posterior odds ratio} = \frac{p_0 f(\theta_0|D)}{(1-p_0)f(\theta_0)},$$

where $f(\theta|D)$ is the posterior distribution when $\theta \neq \theta_0$, i.e.,

$$f(\theta|D) = \frac{p(D|\theta)f(\theta)}{\int_\Theta p(D|\theta)f(\theta)d\theta}.$$

Then the Bayes factor is given by

$$\mathbf{B}_{01} = \frac{f(\theta_0|D)}{f(\theta_0)},$$

which is called the Savage-Dickey density ratio (SDDR).

Predictive Distribution i

Let $\tilde{x}$ denote an unrealized/future value of the population distribution $p(x|\theta)$. Since it is a random variable, we can consider the joint distribution of $\tilde{x}$ and the previous data $D = (x_1, \ldots, x_n)$:

$$p(\tilde{x}, x_1, \ldots, x_n) = p(\tilde{x}, D),$$

Then, from the definition of the conditional probability, we have

$$p(\tilde{x}, D) = p(\tilde{x}|D)p(D) \quad \Rightarrow \quad p(\tilde{x}|D) = \frac{p(\tilde{x}, D)}{p(D)}.$$

Predictive Distribution ii

Furthermore, both $p(D)$ and $p(\tilde{x}, D)$ are regarded as the marginal likelihood given D and $(\tilde{x}, D)$ respectively, that is,

$$p(D) = \int_{\Theta} p(D|\theta)p(\theta)d\theta,$$
$$p(\tilde{x}, D) = \int_{\Theta} p(\tilde{x}, D|\theta)p(\theta)d\theta.$$

In sum, we have

$$p(\tilde{x}|D) = \frac{\int_{\Theta} p(\tilde{x}, D|\theta)p(\theta)d\theta}{\int_{\Theta} p(D|\theta)p(\theta)d\theta}.$$

Predictive Distribution iii

This is called the predictive distribution of $\tilde{x}$. In particular, if $\tilde{x}$ and D are independent, we have

$$p(\tilde{x}, D|\theta) = p(\tilde{x}|\theta)p(D|\theta).$$

Thus⍰the predictive distribution of $\tilde{x}$ is rearranged as

$$\begin{aligned} p(\tilde{x}|D) &= \frac{\int_\Theta p(\tilde{x}|\theta)p(D|\theta)p(\theta)d\theta}{\int_\Theta p(D|\theta)p(\theta)d\theta} \\ &= \int_\Theta p(\tilde{x}|\theta)\frac{p(D|\theta)p(\theta)}{\int_\Theta p(D|\theta)p(\theta)d\theta}d\theta \\ &= \int_\Theta p(\tilde{x}|\theta)p(\theta|D)d\theta. \end{aligned}$$

Predictive Distribution (Bernoulli Model) i

Let us derive the predictive distribution for the Bernoulli distribution.

$$
\begin{aligned}
p(\tilde{x}|D) &= \int_{\Theta} p(\tilde{x}|\theta)p(\theta|D)d\theta \\
&= \int_0^1 \theta^{\tilde{x}}(1-\theta)^{1-\tilde{x}} \frac{\theta^{\alpha_\star-1}(1-\theta)^{\beta_\star-1}}{B(\alpha_\star,\beta_\star)}d\theta \\
&= \frac{\int_0^1 \theta^{\tilde{x}+\alpha_\star-1}(1-\theta)^{\beta_\star-\tilde{x}}d\theta}{B(\alpha_\star,\beta_\star)} \\
&= \frac{B(\alpha_\star+\tilde{x},\beta_\star-\tilde{x}+1)}{B(\alpha_\star,\beta_\star)},
\end{aligned}
$$

Predictive Distribution (Bernoulli Model) ii

Using

$$B(\alpha+1,\beta)=\frac{\alpha}{\alpha+\beta}B(\alpha,\beta),$$
$$B(\alpha,\beta+1)=\frac{\beta}{\alpha+\beta}B(\alpha,\beta),$$

we have

$$p(\tilde{x}=1|D)=\frac{B(\alpha_\star+1,\beta_\star)}{B(\alpha_\star,\beta_\star)}=\frac{\alpha_\star}{\alpha_\star+\beta_\star},$$
$$p(\tilde{x}=0|D)=\frac{B(\alpha_\star,\beta_\star+1)}{B(\alpha_\star,\beta_\star)}=\frac{\beta_\star}{\alpha_\star+\beta_\star}.$$

Predictive Distribution (Bernoulli Model) iii

Finally

$$p(\tilde{x}|D) = \left(\frac{\alpha_\star}{\alpha_\star + \beta_\star}\right)^{\tilde{x}} \left(\frac{\beta_\star}{\alpha_\star + \beta_\star}\right)^{1-\tilde{x}}.$$

This is the Bernoulli distribution with $\theta = \frac{\alpha_\star}{\alpha_\star + \beta_\star}$.

Poisson Distribution

The p.m.f. of a Poisson distribution is

$$p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \ldots, \ \lambda > 0,$$

$$\mathbf{E}[X] = \mathbf{Var}[X] = \lambda.$$

The Poisson distribution is used to model the number of occurrences in a fixed time interval of rare events such as

- traffic accidents
- crimes
- arrival of customers

Gamma Distribution

We use a gamma distribution

$$\boldsymbol{\lambda \sim \mathbf{Gamma}(\alpha_0, \beta_0)},$$

as the prior of $\boldsymbol{\lambda}$ in the Poisson distribution.

The p.d.f. of a gamma distribution $\boldsymbol{\mathbf{Gamma}(\alpha, \beta)}$ is given by

$$p(x|\alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \quad x > 0,\ \alpha > 0,\ \beta > 0,$$

where $\Gamma(\cdot)$ is the gamma function:

$$\Gamma(x) = \int_0^{\infty} z^{x-1} e^{-z} dz.$$

Likelihood

The likelihood of λ given $D = (x_1, \ldots, x_n)$ is

$$\begin{aligned} p(D|\lambda) &= \prod_{i=1}^{n} p(x_i|\lambda) \\ &= \prod_{i=1}^{n} \frac{\lambda_i^x e^{-\lambda}}{x_i!} = \frac{\lambda^{\sum_{i=1}^n x_i} e^{-n\lambda}}{\prod_{i=1}^n x_i!}, \\ &= \frac{\lambda^y e^{-n\lambda}}{\prod_{i=1}^n x_i!}, \quad y = \sum_{i=1}^{n} x_i. \end{aligned}$$

Posterior Distribution

Applying Bayes' theorem, we have

$$
\begin{aligned}
p(\lambda|D) &\propto p(D|\lambda)p(\lambda) \\
&\propto \lambda^{y} e^{-n\lambda} \times \lambda^{\alpha_0 - 1} e^{-\beta_0 \lambda} \\
&\propto \lambda^{y+\alpha_0-1} e^{-(n+\beta_0)\lambda} \\
&\propto \lambda^{\alpha_\star - 1} e^{-\beta_\star \lambda} \\
\alpha_\star = y + \alpha_0, &\quad \beta_\star = n + \beta_0.
\end{aligned}
$$

This is the gamma distribution $\mathbf{Gamma}(\alpha_\star, \beta_\star)$.

Natural Conjugate Prior

Some of you already noticed the following observations.

- In case of the Bernoulli distribution, if we use a beta distribution as the prior for the probability of success, the posterior is also a beta distribution.
- In case of the Poisson distribution, if we use a gamma distribution as the prior for the mean, the posterior is also a gamma distribution.

A class of prior distributions that make the posterior distribution belong to the same family of the prior distribution is called the natural conjugate prior or conjugate prior.

Predictive Distribution

The predictive distribution for the Poisson distribution is

$$p(\tilde{x}|D) = \binom{\tilde{x} + \alpha_\star - 1}{\alpha_\star - 1} \left(\frac{\beta_\star}{\beta_\star + 1} \right)^{\alpha_\star} \left(\frac{1}{\beta_\star + 1} \right)^{\tilde{x}},$$
$$\tilde{x} = 0, 1, 2, \ldots$$

This is the negative-binomial distribution.

Proof i

Following the definition of the predictive distribution, we have

$$\begin{aligned}
p(\tilde{x}|D) &= \int_0^\infty \frac{\lambda^{\tilde{x}} e^{-\lambda}}{\tilde{x}!} \frac{\beta_\star^{\alpha_\star}}{\Gamma(\alpha_\star)} \lambda^{\alpha_\star - 1} e^{-\beta_\star \lambda} d\lambda \\
&= \frac{\beta_\star^{\alpha_\star}}{\tilde{x}!\Gamma(\alpha_\star)} \int_0^\infty \lambda^{\tilde{x}+\alpha_\star-1} e^{-(\beta_\star+1)\lambda} d\lambda \\
&= \frac{\Gamma(\tilde{x}+\alpha_\star)}{\tilde{x}!\Gamma(\alpha_\star)} \frac{\beta_\star^{\alpha_\star}}{(\beta_\star+1)^{\tilde{x}+\alpha_\star}}.
\end{aligned}$$

Proof ii

Since $\Gamma(n) = (n-1)!$, if $\alpha_\star$ is an integer, we obtain

$$\frac{\Gamma(\tilde{x} + \alpha_\star)}{\tilde{x}!\Gamma(\alpha_\star)} = \frac{(\tilde{x} + \alpha_\star - 1)!}{\tilde{x}!(\alpha_\star - 1)!} = \binom{\tilde{x} + \alpha_\star - 1}{\alpha_\star - 1}.$$

Therefore

$$\begin{aligned} p(\tilde{x}|D) &= \binom{\tilde{x} + \alpha_\star - 1}{\alpha_\star - 1} \frac{\beta_\star^{\alpha_\star}}{(\beta_\star + 1)^{\tilde{x} + \alpha_\star}} \\ &= \binom{\tilde{x} + \alpha_\star - 1}{\alpha_\star - 1} \left(\frac{\beta_\star}{\beta_\star + 1}\right)^{\alpha_\star} \left(\frac{1}{\beta_\star + 1}\right)^{\tilde{x}}. \end{aligned}$$

Normal Distribution

The p.d.f. of a normal distribution $\mathbf{Normal}(\mu, \sigma^2)$ is

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right],$$

$$-\infty < x < \infty,\ -\infty < \mu < \infty,\ \sigma^2 > 0,$$

$$\mathbf{E}[X] = \mu, \quad \mathbf{Var}[X] = \sigma^2.$$

- The normal distribution is the mainstay of statistics.
- Many economic data are supposed to follow a normal distribution, though it is not the case for financial data.
- Many sophisticated statistical models are built upon the normal distribution.
- The normal distribution is often a limit of the other distribution — the central limit theorem.

Normal-Inverse-Gamma Distribution i

The natural conjugate prior for (μ, σ^2) is

$$\mu|\sigma^2 \sim \textbf{Normal}\left(\mu_0, \frac{\sigma^2}{n_0}\right),\ \sigma^2 \sim \textbf{Inv.Gamma}\left(\frac{\nu_0}{2}, \frac{\lambda_0}{2}\right),$$

where $\textbf{Inv.Gamma}(\cdot)$ represents the inverse gamma distribution:

$$p(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp\left(-\frac{\beta}{x}\right),\ x > 0,\ \alpha > 0,\ \beta > 0.$$

Note that

$$X \sim \textbf{Gamma}(\alpha,\beta) \quad \Rightarrow \quad \frac{1}{X} \sim \textbf{Inv.Gamma}(\alpha,\beta).$$

Normal-Inverse-Gamma Distribution ii

The joint p.d.f. of the prior distribution is given by

$$\begin{aligned} p(\mu, \sigma^2) &= p(\mu|\sigma^2)p(\sigma^2), \\ p(\mu|\sigma^2) &= \sqrt{\frac{n_0}{2\sigma^2}} \exp\left[-\frac{n_0(\mu - \mu_0)^2}{2\sigma^2}\right], \\ p(\sigma^2) &= \frac{\left(\frac{\lambda_0}{2}\right)^{\frac{\nu_0}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)} (\sigma^2)^{-\left(\frac{\nu_0}{2}+1\right)} \exp\left(-\frac{\lambda_0}{2\sigma^2}\right). \end{aligned}$$

The joint distribution of (μ, σ^2) is often called the normal-inverse-gamma distribution.

Likelihoood i

The likelihood of (μ, σ^2) is

$$
\begin{aligned}
p(D|\mu, \sigma^2) &= \prod_{i=1}^{n} p(x_i|\mu, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right] \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2\sigma^2}\right].
\end{aligned}
$$

Likelihoood ii

Since

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\mu)^2 &= \sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-\mu)^2\\ &= \sum_{i=1}^{n}\left\{(x_i-\bar{x})^2+2(x_i-\bar{x})(\bar{x}-\mu)+(\bar{x}-\mu)^2\right\}\\ &= \sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2,\end{aligned}$$

the likelihood is rewritten as

$$p(D|\mu,\sigma^2)\propto(\sigma^2)^{-\frac{n}{2}}\,\mathbf{exp}\left[-\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2}{2\sigma^2}\right].$$

Posterior Distribution Of (μ, σ^2)

Applying Bayes' theorem, we have

$$
\begin{aligned}
& p(\mu, \sigma^2 | D) \\
& \propto p(D|\mu, \sigma^2) p(\mu|\sigma^2) p(\sigma^2) \\
& \propto (\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2}{2\sigma^2} \right] \\
& \quad \times (\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{n_0(\mu - \mu_0)^2}{2\sigma^2} \right] \times (\sigma^2)^{-(\frac{\nu_0}{2}+1)} \exp\left[-\frac{\lambda_0}{2\sigma^2} \right] \\
& \propto (\sigma^2)^{-\frac{n+\nu_0+3}{2}} \exp\left[-\frac{1}{2\sigma^2} \left\{ \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 \right. \right. \\
& \qquad\qquad\qquad\qquad \left. \left. + n_0(\mu - \mu_0)^2 + \lambda_0 \right\} \right].
\end{aligned}
$$

Completing The Square

"Completing the square" is referred to as a transformation of a quadratic functions:

$$ax^2 + bx + c = a\left(x + \frac{b}{2a}\right)^2 + c - \frac{b^2}{4a}.$$

By completing the square, we have

$$\begin{aligned}
&n(\bar{x} - \mu)^2 + n_0(\mu - \mu_0)^2 \\
&= (n + n_0)\mu^2 - 2(n\bar{x} + n_0\mu_0)\mu + n\bar{x}^2 + n_0\mu_0^2 \\
&= (n + n_0)\left(\mu - \frac{n\bar{x} + n_0\mu_0}{n + n_0}\right)^2 + \frac{nn_0}{n + n_0}(\mu_0 - \bar{x})^2.
\end{aligned}$$

Joint Posterior Distribution i

Therefore the joint posterior distribution of (μ, σ^2) is

$$p(\mu, \sigma^2|D) \propto (\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{n_*(\mu-\mu_*)^2}{2\sigma^2}\right] \\ \times (\sigma^2)^{-\left(\frac{\nu_*}{2}+1\right)} \exp\left(-\frac{\lambda_*}{2\sigma^2}\right),$$

where

$$\mu_* = \frac{n\bar{x}+n_0\mu_0}{n+n_0}, \quad n_* = n+n_0, \quad \nu_* = n+\nu_0,$$

$$\lambda_* = \sum_{i=1}^{n}(x_i-\bar{x})^2 + \frac{nn_0}{n+n_0}(\mu_0-\bar{x})^2 + \lambda_0.$$

Joint Posterior Distribution ii

This is also a normal-inverse-gamma distribution.

$$\begin{aligned} \mu | \sigma^2, D &\sim \mathbf{Normal}\left(\mu_*, \frac{\sigma^2}{n_*}\right), \\ \sigma^2 | D &\sim \mathbf{Inv.Gamma}\left(\frac{\nu_*}{2}, \frac{\lambda_*}{2}\right). \end{aligned}$$

Integrating Out Nuisance Parameters

To proceed to posterior inference on the mean μ, we need the marginal posterior distribution $p(\mu|D)$. This is obtained by integrating out the other parameter σ^2 as

$$\begin{aligned} p(\mu|D) &= \int_0^\infty p(\mu, \sigma^2|D) d\sigma^2 \\ &= \int_0^\infty p(\mu|\sigma^2, D) p(\sigma^2|D) d\sigma^2. \end{aligned}$$

A parameter that is not the primary subject of our analysis is called the nuisance parameter.

Marginal Posterior Distribution Of μ i

The marginal posterior distribution of $\boldsymbol{\mu}$ is derived as

$$\begin{aligned}
p(\mu|D) &= \int_0^\infty p(\mu|\sigma^2, D)p(\sigma^2|D)d\sigma^2 \\
&= \sqrt{\frac{n_*}{2\pi}}\frac{\left(\frac{\lambda_*}{2}\right)^{\frac{\nu_*}{2}}}{\Gamma\left(\frac{\nu_*}{2}\right)}\int_0^\infty (\sigma^2)^{-\left(\frac{\nu_*+1}{2}+1\right)}\exp\left[-\frac{n_*(\mu-\mu_*)^2+\lambda_*}{2\sigma^2}\right]d\sigma^2 \\
&= \frac{\sqrt{n_*}\lambda_*^{\frac{\nu_*}{2}}2^{-\frac{\nu_*+1}{2}}\Gamma\left(\frac{\nu_*+1}{2}\right)}{\Gamma\left(\frac{\nu_*}{2}\right)\sqrt{\pi}}\left[\frac{n_*(\mu-\mu_*)^2+\lambda_*}{2}\right]^{-\frac{\nu_*+1}{2}} \\
&= \frac{\Gamma\left(\frac{\nu_*+1}{2}\right)}{\Gamma\left(\frac{\nu_*}{2}\right)}\sqrt{\frac{n_*}{\pi\lambda_*}}\left[1+\frac{n_*(\mu-\mu_*)^2}{\lambda_*}\right]^{-\frac{\nu_*+1}{2}}.
\end{aligned}$$

Marginal Posterior Distribution Of μ ii

The integral is evaluated with the following formula:

$$\int_0^\infty x^{-(\alpha+1)} e^{-\frac{\beta}{x}} dx = \beta^{-\alpha}\Gamma(\alpha).$$

The marginal posterior distribution $p(\mu|D)$ is a (Student's) t-distribution:

$$\mu|D \sim \mathrm{t}\left(\nu_*, \mu_*, \tau_*^2\right), \quad \tau_*^2 = \frac{\lambda_*}{\nu_* n_*},$$

In general, the p.d.f. of the t-distribution $\mathrm{t}(\nu, \mu, \sigma^2)$ is

$$p(x|\nu, \mu, \sigma^2) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu\sigma^2}}\left[1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right]^{-\frac{\nu+1}{2}}.$$

Student's t-Distribution

The t-distribution in introductory statistics is defined as

$$T = \frac{Z}{\sqrt{V/\nu}},\ Z \sim \mathbf{Normal}(0,1),\ V \sim \chi^2(\nu),\ Z \perp V,$$

where $\chi^2(\nu)$ represents the χ^2-distribution with degree of freedom ν and "$\perp$" implies independence. In our notation, $T \sim \mathrm{t}(\nu, 0, 1)$.

In the same manner, $T \sim \mathrm{t}(\nu, \mu, \sigma^2)$ is defined as

$$T = \mu + \frac{\sigma Z}{\sqrt{V/\nu}},\ Z \sim \mathbf{Normal}(0,1),\ V \sim \chi^2(\nu),\ Z \perp V.$$

Predictive Distribution

The predictive distribution for the normal distribution is

$$\tilde{x}|D \sim \mathrm{t}(\nu_*, \mu_*, \tau_*^2(1 + n_*)).$$

Proof i

We muse evaluate

$$p(\tilde{x}|D) = \int_0^\infty \underbrace{\int_{-\infty}^\infty p(\tilde{x}|\mu, \sigma^2) p(\mu|\sigma^2, D) d\mu}_{g(\sigma^2)} p(\sigma^2|D) d\sigma^2.$$

First we derive the closed form of the integral:

$$\begin{aligned} & g(\sigma^2) \\ & = \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\tilde{x}-\mu)^2}{2\sigma^2}\right] \sqrt{\frac{n_*}{2\pi\sigma^2}} \exp\left[-\frac{n_*(\mu-\mu_*)^2}{2\sigma^2}\right] d\mu \\ & = \frac{\sqrt{n_*}}{2\pi\sigma^2} \int_{-\infty}^\infty \exp\left[-\frac{(\tilde{x}-\mu)^2 + n_*(\mu-\mu_*)^2}{2\sigma^2}\right] d\mu. \end{aligned}$$

Proof ii

Note that, by completing the square, we have

$$
\begin{aligned}
&(\tilde{x} - \mu)^2 + n_*(\mu - \mu_*)^2 \\
&= (1 + n_*)\mu^2 - 2(\tilde{x} + n_*\mu_*)\mu + \tilde{x}^2 + n_*\mu_*^2 \\
&= (1 + n_*)\left(\mu - \frac{\tilde{x} + n_*\mu_*}{1 + n_*}\right)^2 + \frac{n_*}{1 + n_*}(\tilde{x} - \mu_*)^2.
\end{aligned}
$$

Proof iii

Thus

$$
\begin{aligned}
& g(\sigma^2) \\
&= \sqrt{\frac{n_*}{2\pi\sigma^2(1+n_*)}}\,\mathbf{exp}\left[-\frac{n_*(\tilde{x}-\mu_*)^2}{2\sigma^2(1+n_*)}\right] \\
&\quad \times \underbrace{\int_{-\infty}^{\infty}\sqrt{\frac{1+n_*}{2\pi\sigma^2}}\,\mathbf{exp}\left[-\frac{1+n_*}{2\sigma^2}\left(\mu-\frac{\tilde{x}+n_*\mu_*}{1+n_*}\right)^2\right]d\mu}_{1} \\
&= \sqrt{\frac{n_*}{2\pi\sigma^2(1+n_*)}}\,\mathbf{exp}\left[-\frac{n_*(\tilde{x}-\mu_*)^2}{2\sigma^2(1+n_*)}\right].
\end{aligned}
$$

Proof iv

Substituting this result for $g(\sigma^2)$, we have

$$
\begin{aligned}
p(\tilde{x}|D) &= \sqrt{\frac{n_*}{2\pi(1+n_*)}} \frac{\left(\frac{\lambda_*}{2}\right)^{\frac{\nu_*}{2}}}{\Gamma\left(\frac{\nu_*}{2}\right)} \\
&\qquad \times \int_0^\infty (\sigma^2)^{-\left(\frac{\nu_*+1}{2}+1\right)} \exp\left[-\frac{\frac{n_*}{1+n_*}(\tilde{x}-\mu_*)^2+\lambda_*}{2\sigma^2}\right] d\sigma^2 \\
&= \frac{\Gamma\left(\frac{\nu_*+1}{2}\right)}{\Gamma\left(\frac{\nu_*}{2}\right)} \sqrt{\frac{n_*}{\pi\lambda_*(1+n_*)}} \left[1+\frac{n_*(\tilde{x}-\mu_*)^2}{\lambda_*(1+n_*)}\right]^{-\frac{\nu_*+1}{2}} \\
&= \frac{\Gamma\left(\frac{\nu_*+1}{2}\right)}{\Gamma\left(\frac{\nu_*}{2}\right)\sqrt{\pi\nu_*\tau_*^2(1+n_*)}} \left[1+\frac{(\tilde{x}-\mu_*)^2}{\nu_*\tau_*^2(1+n_*)}\right]^{-\frac{\nu_*+1}{2}},
\end{aligned}
$$

where $\tau_*^2 = \frac{\lambda_*}{\nu_* n_*}$.

Summary Of Bayesian Analysis

- $\theta - m \times 1$ vector of unknown parameters
- D — data
- $p(x|\theta)$ — population distribution
- $p(\theta)$ — prior distribution
- $p(D|\theta)$ — likelihood
- $p(\theta|D)$ — posterior distribution

Bayes' Theorem

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int_{\mathbb{R}^m} p(D|\theta)p(\theta)d\theta}.$$

Difficulties In Bayesian Analysis

1. In Bayesian analysis, we are required to evaluate multiple integrals to proceed statistical inference.
2. For example, we need $\int_{\mathbb{R}^m} p(D|\theta)p(\theta)d\theta$ to derive the exact expression of the posterior distribution, but it is not available except for a limited number of examples (e.g., natural conjugate prior).
3. It is the case for the following quantities:
    - Marginal distribution: $p(\theta_j|D) = \int_{\mathbb{R}^{m-1}} p(\theta|D)d\theta_{-j}$.
    - Mean: $\mathbf{E}_{\theta}[\theta_j|D] = \int_{-\infty}^{\infty} \theta_j p(\theta_j|D)d\theta_j$.
    - Median: $\int_{-\infty}^{\text{Median}_{\theta_j}} p(\theta_j|D)d\theta_j = \frac{1}{2}$.
    - Probability: $\mathbf{P}(a \leqq \theta_j \leqq b|D) = \int_a^b p(\theta_j|D)d\theta_j$.
    - Predictive dist.: $p(\tilde{x}|D) = \int_{\mathbb{R}^m} p(\tilde{x}|\theta)p(\theta|D)d\theta$.

MAP Estimation

One exception is the posterior mode:

$$\text{Mode}_{\boldsymbol{\theta}} = \operatorname*{\mathbf{arg\,max}}_{\boldsymbol{\theta} \in \mathbb{R}^m} p(\boldsymbol{\theta}|D).$$

Note that the posterior mode is equivalent to

$$\text{MAP}_{\boldsymbol{\theta}} = \operatorname*{\mathbf{arg\,max}}_{\boldsymbol{\theta} \in \mathbb{R}^m} p(D|\boldsymbol{\theta})p(\boldsymbol{\theta}),$$

which does not depend on the normalizing constant. Since the exact expressions of the prior distribution and the likelihood are known in many applications, we can solve the above optimization problem without evaluating the normalizing constant.

The posterior mode is often called MAP (maximum a posteriori) estimate.

Monte Carlo Methods i

Monte Carlo methods are widely applied numerical techniques to evaluate integrals. Suppose we need to evaluate the following expectation:

$$\mathbf{E}_\theta[h(\theta)] = \int_{\mathbb{R}^m} h(\theta) p(\theta|D) d\theta.$$

Examples:

- Mean $\mathbf{E}_\theta[\theta_j|D]$: $h(\theta) = \theta_j$.
- Probability $\mathbf{P}(a \leqq \theta_j \leqq b|D)$: $h(\theta) = 1_{[a,b]}(\theta_j)$.
- Predictive distribution $p(\tilde{x}|D)$: $h(\theta) = p(\tilde{x}|\theta)$.

Monte Carlo Methods ii

Suppose we have a set of pseudo-random numbers generated from the posterior distribution $p(\theta|D)$, $\{\theta^{(1)}, \dots, \theta^{(T)}\}$, which is called the Monte Carlo sample, and define

$$\hat{h} = \frac{1}{T} \sum_{t=1}^{T} h(\theta^{(t)}).$$

If the law of large numbers holds,

$$\hat{h} \overset{a.s.}{\to} \mathbf{E}_{\theta}[h(\theta)] \quad \text{as} \quad T \to \infty,$$

where $\overset{a.s.}{\to}$ means almost surely convergence.

Monte Carlo Methods iii

Therefore, when T is sufficiently large, $\mathbf{E}_{\theta}[h(\theta)]$ can be well approximated with $\hat{h}$. This is the basic idea behind the Monte Carlo integration method.

Monte Carlo approximations of the posterior statistics are given as follows:

- Mean: $\hat{\mathbf{E}}_{\theta}[\theta_j|D] = \frac{1}{T}\sum_{t=1}^{T}\theta_j^{(t)}$.
- Probability: $\widehat{\mathbf{P}}(a \leqq \theta_j \leqq b|D) = \frac{1}{T}\sum_{t=1}^{T}1_{[a,b]}(\theta_j^{(t)})$.
- Predictive distribution: $\hat{p}(\tilde{x}|D) = \frac{1}{T}\sum_{t=1}^{T}p(\tilde{x}|\theta^{(t)})$.

Monte Carlo Methods iv

Note that the Monte Carlo method merely gives us an approximated value of the true posterior statistic (the true one is obtained when $T \longrightarrow \infty$). So it is contaminated with numerical errors. These errors are measured by

Numerical errors in Monte Carlo approximation

$$\mathbf{SE}[\hat{h}] = \sqrt{\frac{1}{T(T-1)} \sum_{t=1}^{T} \left(h(\theta^{(t)}) - \hat{h}\right)^2}.$$

Kernel Density Estimation

The marginal posterior p.d.f. $p(\theta_j|D)$ can be evaluated with a kernel density estimation method:

$$\hat{p}(\theta_j) = \frac{1}{Tw}\sum_{t=1}^{T} K\left(\frac{\theta_j - \theta_j^{(t)}}{w}\right),$$

where $K(\cdot)$ is a kernel function. In practice, the p.d.f. of the standard normal distribution:

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

is often used as $K(\cdot)$, which is called the Gaussian kernel. w is the bandwidth and determines the smoothness of the evaluated density function.

Monte Carlo Approximation Of Quantiles i

Quantile

The $100\boldsymbol{q}$% quantile $\boldsymbol{\theta}_j^{[q]}$ is defined as

$$\mathbf{P}(\theta_j \leqq \theta_j^{[q]} | D) = q, \quad 0 < q < 1.$$

We can approximate the quantile $\boldsymbol{\theta}_j^{[q]}$ with

Sample Quantile

$$\hat{\theta}_j^{[q]} = \max_{1 \leqq t \leqq T} \theta_j^{(t)} \quad \text{s.t.} \quad \frac{1}{T} \sum_{s=1}^{T} 1_{\left(-\infty,\ \theta_j^{(t)}\right)} \left(\theta_j^{(s)}\right) \leqq q,$$

Monte Carlo Approximation Of Quantiles ii

For example, when $T = 10,000$, $\hat{\theta}_j^{[0.05]}$ is the 500th smallest value in the Monte Carlo sample.

- Median: $\hat{\theta}_j^{[0.5]}$.
- $100(1-q)$% credible interval: $\hat{\theta}_j^{[\frac{q}{2}]} \leqq \theta_j \leqq \hat{\theta}_j^{[1-\frac{q}{2}]}$.
- $100(1-q)$% HPDI [Chen and Shao (1999)]:

$$\hat{\theta}_j^{[\frac{t^*}{T}]} \leqq \theta_j \leqq \hat{\theta}_j^{[1-q+\frac{t^*}{T}]}, \quad 1 \leqq t \leqq qT,$$

where

$$t^* = \arg\min_{1 \leqq t \leqq qT} \left| \hat{\theta}_j^{[\frac{t}{T}]} - \hat{\theta}_j^{[1-q+\frac{t}{T}]} \right|.$$

Markov Chain Monte Carlo (MCMC)

- So far we suppose we have the Monte Carlo sample generated from the posterior distribution.
- However it is not obvious how to generate random numbers from an unknown distribution.
- For this purpose, we apply Markov chain sampling methods (e.g., Gibbs sampler, Metropolis-Hastings algorithm, Hamiltonian Monte Carlo method).
- Markov chain sampling + Monte Carlo integration = Markov chain Monte Carlo (MCMC)
- I will not cover these methods because they are so complicated. Please read books in the reading list.
- Instead we use PyMC for MCMC implementation.

Convergence Diagnostics

Unfortunately Markov chain sampling does not always work, even though the distribution of generated random numbers must converge to the posterior distribution in theory.

To make it work properly, we need careful tuning and monitoring. PyMC can automatically tune the sampling algorithm to some extent. So we discuss how to check the convergence of generated chains.

- n — the number of draws
- m — the number of chains
- θ_{ij} — draw i in chain j $(i = 1, \ldots, n,\ j = 1, \ldots, m)$
- $\{\theta_1, \ldots, \theta_T\}$ — Monte Carlo sample $(T = m \times n)$

Numerical Inefficiency

Effective sample size

$$\hat{T}_e = \frac{T}{1 + 2\sum_{s=1}^{S} \hat{\rho}_s},$$

where

$$\hat{\rho}_s = \frac{\hat{\gamma}_s}{\hat{\gamma}_0},\ \hat{\gamma}_s = \frac{1}{T}\sum_{t=s+1}^{T} (\theta_t - \bar{\theta})(\theta_{t-s} - \bar{\theta}),\ \bar{\theta} = \frac{1}{T}\sum_{t=1}^{T} \theta_t.$$

$\hat{\rho}_s$ is called the autocorrelation (of lag s). If a chain of random numbers are independent, all autocorrelations are zeros. In practice, we often observe strong and persistent positive autocorrelations in the chain, which leads to a smaller $\hat{T}_e$.

Gelman-Rubin Convergence Diagnostic

Gelman-Rubin convergence diagnostic $\hat{R}$

$$\hat{R} = \sqrt{\frac{\hat{V}}{\hat{W}}}, \quad \hat{V} = \frac{n-1}{n}\hat{W} + \frac{1}{n}\hat{B},$$

$$\hat{B} = \frac{n}{m-1}\sum_{j=1}^{m}\left(\bar{\theta}_{.j} - \bar{\theta}\right)^2, \quad \bar{\theta}_{.j} = \frac{1}{n}\sum_{i=1}^{n}\theta_{ij},$$

$$\hat{W} = \frac{1}{m}\sum_{j=1}^{m}s_j^2, \quad s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\theta_{ij} - \bar{\theta}_{.j}\right)^2.$$

If the mean of each chain is equal to each other, $\frac{1}{n}\hat{B}$ will be close to zero when n is sufficiently large. Therefore $\hat{R}$ must be close to one; otherwise, the convergence is doubtful.

Relationship between Wage and Schooling

In econometrics, we formulate relationship between wage and schooling as a mathematical formula. For example, we may suppose their relationship is linear, i.e.,

$$\text{Wage} = \text{Intercept} + \text{Slope} \times \text{Years of Schooling},$$

or

$$\text{Wage} = \alpha + \beta \times \text{Years of Schooling}.$$

By convention, the intercept and the slope are denoted by α and β respectively.

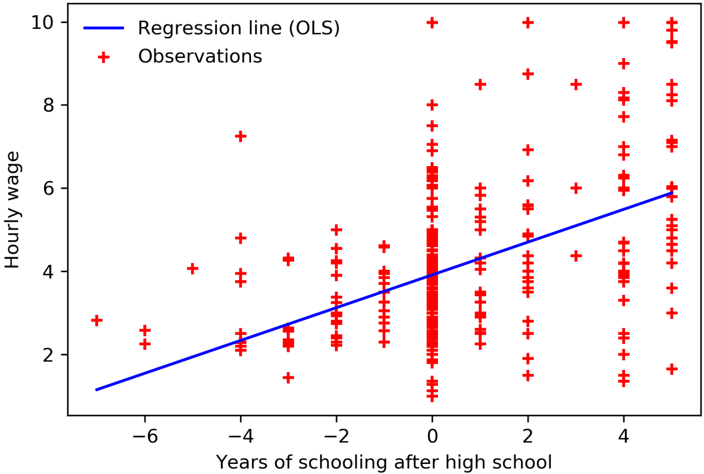

Figure 8: Years of Schooling and Wage

Regression Model

- As we see in the previous figure, not all data points are on the straight line. Moreover, it is impossible to put all of them on the same straight line or curve.
- To accommodate the discrepancy between the straight line and the actual data points, we need to add error or noise to the linear function, that is,

$$\text{Wage} = \alpha + \beta \times \text{Years of Schooling} + \text{Error}.$$

- Such a linear function with additive error is called a (simple linear) regression model.
- The error term is supposed to be random.
- In this lecture, we assume the error term follows a normal distribution.

Model Setup

Suppose we have data $(y_1, x_1), \ldots, (y_n, x_n)$. Let u_i denote the error term for the i-th pair (y_i, x_i), $(i = 1, \ldots, n)$. We assume each u_i independently follows $\mathbf{Normal}(0, \sigma^2)$. Then the regression model is given by

$$y_i = \alpha + \beta x_i + u_i, \quad u_i \sim \mathbf{Normal}(0, \sigma^2),$$

where α, β and σ^2 are unknown parameters.

y_i	x_i	u_i
dependent variable	independent variable	error term
explained variable	explanatory variable	disturbance term
regressand	regressor	noise
response	covariate	innovation

Multiple Regression Model i

In general, a regression model may include more than one independent variable. Suppose the conditional expectation of y_i is a linear combination of k independent variables $(x_{1i}, \ldots, x_{ki})$, i.e.,

$$\mathbf{E}[y_i|x_i] = \mathbf{E}[y_i|x_{1i}, \ldots, x_{ki}] = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}.$$

In practice, we almost always put the constant term β_0 in the model. By adding the error term, we have

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + u_i,\ u_i \sim \mathbf{Normal}(0, \sigma^2).$$

This is sometimes called the multiple regression model.

Multiple Regression Model ii

Now define

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{ki} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix}, u = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}.$$

Then the regression model is summarized as

$$y = X\beta + u, \quad u \sim \mathbf{Normal}(0, \sigma^2 \mathrm{I}).$$

A typical prior distribution for (β, σ^2) is

$$\beta \sim \mathbf{Normal}(\mu_\beta, \Omega_\beta), \quad \sigma^2 \sim \mathbf{Inv.Gamma}\left(\frac{\nu_0}{2}, \frac{\lambda_0}{2}\right).$$

Non-Standard Dependent Variables

In the regression model,

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + u_i,$$

we implicitly suppose the dependent variable y_i in the regression model is continuous and can be either positive or negative. In other words, y_i is supposed to be a real-valued random variable.

As long as the normality assumption is valid and the conditional expectation of y_i is expressed as a linear function:

$$\mathbf{E}[y_i|x_i] = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki},$$

this assumption seems benign. In some cases, however, it is impractical to assume so.

Binary Data

Suppose y_i takes either 1 or 0 with the constant probability, i.e.,

$$y_i = \begin{cases} 1, & \text{with probability } q_i; \\ 0, & \text{with probability } 1 - q_i, \end{cases}$$

which is called a Bernoulli distribution. This type of data appears in empirical analysis on decision making (e.g., consumer's choice) or events (e.g., bankruptcy).

In this case, the conditional expectation of y_i is q_i, the conditional probability $\mathbf{P}(y_i = 1|x_i)$ itself. Since $0 \leqq q_i \leqq 1$ by definition, it is unrealistic to assume that q_i is a linear function of y_i.

Count Data

Suppose y_i takes non-negative integers, $0, 1, 2, \ldots$. This type of data is used in analyzing occurrences of rare events such as traffic accidents, crimes, mechanical failures, and so forth.

It is often assume that y_i follows a Poisson distribution:

$$\mathbf{P}(y_i = y|x_i) = \frac{e^{-\lambda_i}\lambda_i^y}{y!},\ y = 0, 1, 2, \ldots,\ \lambda_i > 0.$$

In this case, the conditional expectation of y_i is λ_i, which must be positive. Thus the linearity assumption on the conditional expectation of y_i seems inappropriate.

Generalized Linear Model

Let $\boldsymbol{\mu}_i$ denote the conditional expectation $\mathbf{E}[y_i|x_i]$ and

$$x_i^\top \beta = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki},$$

for brevity. To relax the limitation of the linearity assumption $\boldsymbol{\mu}_i = x_i^\top \beta$, we introduce a transformation of the conditional expectation $g(\cdot)$:

$$g(\mu_i) = x_i^\top \beta \quad \text{or} \quad \mu_i = g^{-1}(x_i^\top \beta).$$

$g(\cdot)$ is called a link function and a regression-type model of the transformed conditional expectation is called a generalized linear model (GLM).

Link Functions

1. Logit link

$$\log \frac{\mu_i}{1-\mu_i} = x_i^\top \beta \quad \Leftrightarrow \quad \mu_i = \frac{1}{1+e^{-x_i^\top \beta}}.$$

2. Probit link

$$\Phi^{-1}(\mu_i) = x_i^\top \beta \quad \Leftrightarrow \quad \mu_i = \Phi(x_i^\top \beta),$$

where $\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz.$

3. Log link

$$\log \mu_i = x_i^\top \beta \quad \Leftrightarrow \quad \mu_i = e^{x_i^\top \beta}.$$

Remarks

1. Since $\boldsymbol{\mu}_i = x_i^\top \beta$ in the linear regression model, $g(\boldsymbol{\mu}_i) = \boldsymbol{\mu}_i$, which is called the linear link function.
2. Since both logit and probit link functions assure that $\boldsymbol{\mu}_i$ takes a value between 0 and 1, they are suitable for the binary data model.
3. When the probability $\boldsymbol{p}_i$ is transformed with the logit link function, such a GLM is called a logit model.
4. When the probability $\boldsymbol{p}_i$ is transformed with the probit link function, such a GLM is called a probit model.
5. Since the log link function assures that $\boldsymbol{\mu}_i$ is positive, it is used in the Poisson model of count data.

Binary Choice Model

A GLM of Bernoulli binary data with either logit or probit link function is often called a binary choice model, though the binary data are not necessarily related to decision making.

Examples

- Consumer's choice
  $y_i = 1,$ if Consumer i owns an iPhone$;\ 0,$ otherwise.
- Labor force participation
  $y_i = 1,$ if Person i works$;\ 0,$ otherwise.
- Bankruptcy
  $y_i = 1,$ if Firm i goes bankrupt$;\ 0,$ otherwise.

Likelihood in the Logit / Probit Model

Since the probability of y_i is expressed as

$$\mathbf{P}(y_i = y|x_i) = q_i^y(1-q_i)^{1-y},\ q_i = g^{-1}(x_i^\intercal\beta),\ y = 0,1,$$

where g is either logit or probit link function. The joint probability of $D = (y_1,\dots,y_n)$ is given by

$$p(D|\beta) = \prod_{i=1}^{n} q_i^{y_i}(1-q_i)^{1-y_i}.$$

This is the likelihood in the logit / probit model. Since this likelihood is a complicated non-linear function of β, it is impractical to derive the closed form of the posterior distribution.

Poisson Regression Model

A GLM with Poisson count data with the log link function is called a Poisson regression model. Since the probability of y_i is expressed as

$$\mathbf{P}(y_i = y|x_i) = \frac{\lambda_i^y e^{-\lambda_i}}{y!}, \ \lambda_i = e^{x_i^\intercal \beta}, \ y = 0, 1, 2, \ldots,$$

the likelihood is given by

$$p(D|\beta) = \prod_{i=1}^{n} \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \propto \prod_{i=1}^{n} \lambda_i^{y_i} \cdot e^{-\sum_{i=1}^{n} \lambda_i}.$$

Since the above likelihood is a complicated non-linear function of β, it is impractical to derive the closed form of the posterior distribution.