Methods for Statistical Prediction
Financial Time Series I

Topic 2: Methods of Estimation

Hung Chen
Department of Mathematics
National Taiwan University
9/28/2002

OUTLINE

1. Motivated Examples
2. Statistical Models
3. Maximum Likelihood Estimates
4. Substitution Principles
   - Method of Moments
   - Frequency Substitution
5. Method of Least Squares

Motivated Examples

Example 1. Censored exponentially distributed survival times

- Suppose W is a nonnegative random variable having an exponential distribution with mean $\theta > 0$. Its pdf is given by

  $$f_W(w;\theta) = \theta^{-1}\exp(-w/\theta)I_{(0,\infty)}(w),$$

  where the indicator function $I_{(0,\infty)}(w) = 1$ for $w > 0$ and is zero elsewhere. The distribution function is given by

  $$F_W(w;\theta) = \{1 - \exp(-w/\theta)\}I_{(0,\infty)}(w).$$

- In survival or reliability analyses, a study to observe a random variable $W_1, \ldots, W_n$ will generally be terminated in practice before all of these random variables are able to be observed.

  – Let $\mathbf{y} = (y_1^T, \ldots, y_n^T)^T$ denote the observed data, where $y_j = (c_j, \delta_j)^T$ and $\delta_j = 0$ or 1 according as the observation W_j is censored or uncensored at c_j $(j = 1, \ldots, n)$.

- If the observation W_j is uncensored, its realized value w_j is equal to c_j.
- If the observation W_j is censored at c_j, then w_j is some value greater than c_j $(j = 1, \ldots, n)$.
- In medical study, it is commonly assumed that the censored data is caused by competing risk.

  Under this assumption, it is assumed that $(W_1, R_1), \ldots, (W_n, R_n)$ are iid, $C_i = \min(W_i, R_i)$, and $\delta_i = 1_{\{W_i \le R_i\}}$.

  Here R is a nonnegative random variable.

- Approach 1: Model C directly.
  - Derive the density function of C. Note that

    $$\begin{aligned} P(C \le y) &= P(\min(W, R) \le y) \\ &= 1 - P(W > y, R > y) = 1 - e^{-y/\theta}[1 - F_R(y)] \end{aligned}$$

    and hence

    $$f_C(c) = \theta^{-1} e^{-c/\theta}[1 - F_R(c)] + e^{-c/\theta} f_R(c).$$

  - Assuming R is exponentially distributed with mean λ, we have

$$f_C(c) = \frac{\theta+\lambda}{\theta\lambda} \exp\left(-\frac{\theta+\lambda}{\theta\lambda} c\right).$$

Hence, C is again exponentially distributed with mean $\theta\lambda/(\theta+\lambda)$.

- How do we estimate θ?
  We should use information contained in δ_j. Note that δ is a Bernoulli random variable with probability of success

  $$\begin{aligned} P(W \le R) &= \int_0^\infty \int_0^r \theta^{-1} e^{-y/\theta} f_R(r) dy dr \\ &= \int_0^\infty f_R(r)[1 - \exp(-r/\theta)]dr \\ &= 1 - \int_0^\infty e^{-r/\theta} f_R(r) dr. \end{aligned}$$

  When R is exponentially distributed with mean λ, we have

  $$P(W \le R) = \frac{\lambda}{\lambda+\theta}.$$

  By the law of large numbers, we consider using $n^{-1} \Sigma_i \delta_i$ to estimate $P(W \le R)$.

- Approach II: Method of Maximum Likelihood
  - We have iid observations $(C_1, \delta_1), \ldots, (C_n, \delta_n)$ and need to find the probability density

function of (C, δ). Observe that

$$\begin{aligned} P(C \leq c, \delta = 1) &= P(W \leq R, W \leq c) \\ &= \int_0^c \int_0^r f_W(w) f_R(r) dw dr + \int_c^\infty \int_0^c f_W(w) f_R(r) dw dr \\ &= \int_0^c F_W(r) f_R(r) dr + \int_c^\infty F_W(c) f_R(r) dr \\ &= \int_0^c F_W(r) f_R(r) dr + F_W(c)[1 - F_R(c)]. \end{aligned}$$

Then

$$f(C = c, \delta = 1) = f_W(c)[1 - F_R(c)].$$

By the same argument, we have

$$f(C = c, \delta = 0) = [1 - F_W(c)] f_R(c).$$

– The likelihood function is

$$\begin{aligned} &\prod_i (f_W(w_i)\,[1 - F_R(w_i)])^{\delta_i} (f_R(w_i)\,[1 - F_W(w_i)])^{1-\delta_i} \\ &\quad = \prod_i (f_W(w_i))^{\delta_i} \left[1 - F_W(w_i)\right]^{1-\delta_i} \\ &\qquad \cdot \prod_i (f_R(w_i))^{1-\delta_i} \left[1 - F_R(w_i)\right]^{\delta_i}. \end{aligned}$$

– For simplicity, we relabel the observations such that $W_1, \ldots, W_r$ denote the r uncensored observations and $W_{r+1}, \ldots, W_n$ the $n - r$ censored observations.
The likelihood function for θ formed on

the basis of $\mathbf{y}$ is given by

$$\prod_{i=1}^{r}[\theta^{-1}\exp(-w_i/\theta)]\prod_{i=r+1}^{n}\{1-[1-\exp(-w_i/\theta)]\}$$
$$=\theta^{-r}\exp(-\sum_{i=1}^{n}c_i/\theta).$$

– In this case, the MLE of θ can be derived explicitly from the standard differentiation technique.

$$\hat{\theta}=\sum_{i=1}^{n}c_i/r.$$

– Rewrite $\hat{\theta}$ as

$$\left[n^{-1}\sum_{i=1}^{n}c_i\right]/(r/n).$$

It can be shown that $\hat{\theta}$ will converge to θ in probability.

Remarks:

- The exponential distribution is often used to model lifetimes or waiting times.

- Suppose that we consider modeling the lifetime of an electronic component, T, as an exponential random variable with parameter θ. Its implication is as follows:

$$P(T>t+s|T>s)=\frac{P(T>t+s\text{ and }T>s)}{P(T>s)}$$

$$
\begin{aligned}
&= \frac{P(T > t + s)}{P(T > s)} = \frac{e^{-(t+s)/\theta}}{e^{-s/\theta}} \\
&= \exp(-t/\theta).
\end{aligned}
$$

This is so-called memoryless property of exponential distribution.

- Does it make sense to use exponential distribution to model human lifetimes? Compare the probability that a 16-year-old will live at least 10 more years and the probability that a 80-year-old will live at least 10 more years.

- **Hazard function** $h(t)$: It is defined as
$$h(t) = \frac{f(t)}{1 - F(t)} (= -\frac{d}{dt} \log S(t)),$$
where $S(t) = P(T > t) = 1 - F(t)$. I can be thought of as the instantaneous death rate for individuals who are alive at time t. If an individual is alive at time t, the probability that that individual will die in the time interval $(t, t + \delta)$ is, assuming that the density function is continuous at t,
$$P(t \leq T \leq t + \delta | T \geq t) = \frac{P(t \leq T \leq t + \delta)}{P(T \geq t)}$$

$$= \frac{F(T \leq t + \delta) - F(t)}{1 - F(t)} \approx \frac{\delta f(t)}{1 - F(t)}.$$

- For an exponential random variable T with mean θ, its hazard function is $1/\theta$ (a constant function).
  As a remark, the expectation of an exponential random variable is θ.
  Do you think that the connection between the expectation and the hazard function is a coincidence? Is there any intuitive explanation?

- Usually, the hazard function of human lifetimes is assumed to be of bathtub shape.
  How would you model the density function of human lifetimes?

Example 2. Model heterogeneous data by finite-mixture models

- In the problem considered by Do and McLachlan (1984), the population of interest consists of rats from g species $G_1, \dots, G_g$, that are consumed by owls in some unknown proportions $\pi_1, \dots, \pi_g$.

- The problem is to estimate the π on the basis of the observation vector $\mathbf{W}$ containing measurements recorded on a sample of size n of rat skulls taken from owl pellets.
  The rats constitute part of an owl's diet, and indigestible material is regurgitated as a pellet.

- Use the argument of conditioning, the underlying population can be modeled as consisting of g distinct groups $G_1, \dots, G_g$ in some unknown proportions $\pi_1, \dots, \pi_g$, and where the conditional pdf of $\mathbf{W}$ given membership of the ith group G_i is $f_i(\mathbf{w})$.

- Let $\mathbf{y} = (w_1^T, \dots, w_n^T)^T$ denote the observed random sample obtained from the mixture density

$$f(w; (\pi_1, \ldots, \pi_{g-1})) = \sum_{i=1}^{g} \pi f_i(w).$$

- The log likelihood function for $(\pi_1, \ldots, \pi_{g-1})$ can be formed from the observed data $\mathbf{y}$ is given by

$$\sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{g} \pi_j f_j(w_i) \right\}.$$

- On differentiating log likelihood function with respect to π_j $(j = 1, \ldots, g-1)$, we obtain

$$\sum_{i=1}^{n} \left\{ \frac{f_j(w_i)}{f(w_i; (\pi_1, \ldots, \pi_{g-1}))} - \frac{f_g(w_i)}{f(w_i; (\pi_1, \ldots, \pi_{g-1}))} \right\} = 0,$$

for $j = 1, \ldots, g-1$. It clearly does not yield an explicit solution for $(\pi_1, \ldots, \pi_{g-1})^T$.

Statistical models

- Most studies and experiments, scientific or industrial, large scale or small, produce data whose analysis is the ultimate object of the endeavor.
  - Compare the efficiency of two ways of doing something under similar conditions such as: brewing coffee; reducing pollution; treating a disease; producing energy; learning a maze; and so on.
  - Abstraction: It can be thought of as a problem of comparing the efficacy of two methods applied to the members of a certain population.
  - Run $m + n$ independent experiments as follows: $m + n$ members of the population are picked at random and m of these are assigned to the first method and the remaining n are assigned to the second method.
  - In comparing two drugs A and B we would administer drug A to m and drug B to n randomly selected patients and then

measure temperature, blood pressure, have the patient rated quantitatively for improvement by physicians, and so on.

- Random variability would come primarily from differing responses among patients to the same drug, but also from error in the measurements and variation in the purity of the drugs.
- one sample location model for measurement:
  Let $X_1, \ldots, X_n$ be the n determinations of μ. Write
  $$X_i = \mu + \epsilon_i, \quad 1 \le i \le n,$$
  where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ is the vector of errors.
- two-sample problem:
  Let $X_1, \ldots, X_n$ be the n samples from the population with distribution F and $Y_1, \ldots, Y_m$ be the m samples from the population with distribution G.

- Many statistical procedures are based on statistical models which specify under which conditions the data are generated.

- Usually the assumption is made that the set of observations $x_1, \ldots, x_n$ is a set of (i) independent random variables (ii) identically distributed with common pdf $f(x_i, \boldsymbol{\theta})$.

- Once this model is specified, the statistician tries to find optimal solutions to his problem (usually related to the inference on a set of parameters $\boldsymbol{\theta} \in \Theta \subset R^k$, characterizing the uncertainty about the model).
  Does this statement fit to the just-mentioned two-sample problem?

- Any statistical inference starts from a basic family of probability measures, expressing our prior knowledge about the nature of the probability measures from where the observations originate.
  A model P is a collection of probability measures P on $(\mathcal{X}, \mathcal{A})$ where $\mathcal{X}$ is the sample space with a σ-field of subsets $\mathcal{A}$.

- If $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, $\Theta \subset R^k$ for some k, then P is a *parametric model.*

- **Example 3.** Bernoulli trials

  - Consider a new model of automobile which is being produced in large numbers.
  - Choose one at random from the production line and observe whether or not it suffers a mechanical breakdown within two years.
  - In each trial, there are only two possible observations. The sample space consists of two elements 1 (representing breakdown) and 0 (representing no breakdown).
  - The inherent variability in the situation is described by a probability distribution which in this case is defined by a single number θ, the probability of breakdown.
  - The possible probability distribution on the sample space can be described by a Bernoulli trial with an unknown parameter θ between 0 and 1.

- **Example 4.** The parameter is a function.
  - Suppose we have a large batch of seeds stored under constant conditions of temperature and humidity.
  - In the course of time seeds die.

Suppose that at time t a proportion $\pi(t)$ of the stored seeds are still alive.

- At each of times $t_1, t_2, \ldots, t_s$ we take a random sample of n seeds and observe how many are still alive.
- A typical observation consists of an ordered set $(r_1, r_2, \ldots, r_s)$ of integers, r_i being the number of seeds observed to be alive at time t_i.
- The appropriate distribution for describing the variable element in this situation is

$$p(r_1, r_2, \ldots, r_s) = \prod_{i=1}^{s} C(n, r_i)[\pi(t_i)]^{r_i}[1-\pi(t_i)]^{n-r_i}.$$

Here $\pi(\cdot)$ is an unknown distribution. In this example, the parameter is a function.

- Isotonic regression problem: $\pi(t)$ is necessarily a non-increasing function of t, taking values between 0 and 1.

Can we find a parametric model for $\pi(t)$?

Related Issues:

- Suppose that we have a fully specified parametric family of models. Denote the parameter of interest by $\boldsymbol{\theta}$.
- Suppose that we wish to calculate from the data a single value representing the "best estimate" that we can make of the unknown parameter. We call such a problem one of *point estimation*.
- Instead of point estimation, we can estimate the parameter by giving a confidence interval which is associated with the probability of covering the true value.

  When we say that a 95% CI of θ is $(0.3, 0.7)$, it does not mean that there is a 95% probability of $\theta \in (0.3, 0.7)$.

  Such a claim does not make any sense since

  – Although θ is unknown, it is still a **fixed** number.

  – $(0.3, 0.7)$ is a known fixed interval.

  – θ is either in $(0.3, 0.7)$ or not in that interval. It will not be sometimes in $(0.3, 0.7)$ or sometimes not in.

  - The precise meaning of probability 0.95 will be discussed later on.
  - The probability 0.95 refers to the probability that θ is in a random interval. Here $(0.3, 0.7)$ is one realization of that random interval.

- Distinction between data and random variables:
  In statistics, we deal with data only.
  Why do we need to introduce random variables?

Attitudes on Models:

- The statistician may be a "pessimist" who does not believe in any particular model $f(x, \boldsymbol{\theta})$. In this case he must be satisfied with descriptive methods (like exploratory data analysis) without the possibility of inductive inference.

- The statistician may be an "optimist" who strongly believes in one model. In this case the analysis is straightforward and optimal solutions may often be easily obtained.

- The statistician may be “realist”: he would like to specify a particular model $f(x, \boldsymbol{\theta})$ in order to get operational results but he may have either some doubt about the validity of this hypothesis or some difficulty in choosing a particular parametric family.

Let us illustrate this kind of preoccupation with an example.

- Suppose that the parameter of interest is the “center” of some population.
- In many situations, the statistician may argue that, due to a central limit effect, the data are generated by a normal pdf.
- In this case the problem is restricted to the problem of inference on μ, the mean of the population.
- But in some cases, he may have some doubt about these central limit effects and may suspect some skewness and/or some kurtosis or he may suspect that some observations are generated by other models (leading to the presence of outliers).

In this context three types of question may be raised to avoid gross errors in the prediction, or in the inference:

- Does the optimal solution, computed for assumed model $f(x, \boldsymbol{\theta})$, still have “good” properties if the true model is a little different?
  This question is concerned with the sensitivity of a given criterion to the hypotheses (criterion robustness).
  Question: Validity of one-sample t-test
  Partial Answer: Central Limit Theorem
- Are the optimal solutions computed for other models near to the original one really substantially different?
  In this question, it is the sensitivity of the inference that is analyzed (inference robustness).

Maximum Likelihood Estimates

- The true distribution on the sample space can be labeled by a parameter $\boldsymbol{\theta}$ taking values in a finite-dimensional Euclidean space.
- We further assume the family $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ $(\Theta \subset R^k)$ possesses density functions $\{p_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ with respect to some natural measure on the sample space, such as counting measure if the sample space is discrete or Lebesgue measure when it is not.
  - In the discrete case, $p_{\boldsymbol{\theta}}(x)$ is the probability of the point x when $\boldsymbol{\theta}$ is the true parameter.
  - In the continuous case, $p_{\boldsymbol{\theta}}(x)$ is the probability density at x when $\boldsymbol{\theta}$ is the true parameter.
- $\mathbf{x}$: the observed set of values obtained in an experiment.
- Consider $p(\mathbf{x}, \boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$ for fixed $\mathbf{x}$.
  $p(\mathbf{x}, \boldsymbol{\theta})$ is called the likelihood function.
  We also write it $L(\boldsymbol{\theta}, \mathbf{x})$.

$L(\boldsymbol{\theta}, \mathbf{x})$ gives the probability of observing $\mathbf{x}$ for each $\boldsymbol{\theta}$ when $\mathbf{X}$ is discrete.

- Idea: Find the value $\hat{\boldsymbol{\theta}}$ of the parameter which is most plausible after we have observed the data.
- A maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ is any element of Θ such that

$$p(\mathbf{x}, \boldsymbol{\theta}(\mathbf{x})) = \max_{\boldsymbol{\theta} \in \Theta} p(\mathbf{x}, \boldsymbol{\theta}).$$

- This principle was first put forward as a novel and original method of deriving estimators by R.A. Fisher in 1922. It very soon proved to be a fertile approach to statistical inference in general, and was widely adopted; but the exact properties of the ensuring estimators and test procedures were only gradually discovered.
- How do we find the maximum of $L(\theta, \mathbf{x})$?
  - A systematic way we learn in calculus is to transform a maximization problem to a root-finding problem.
  - The above strategy may not always work. Refer to the the *uniform* example.

- Computationally feasibility:
  Before 1960, we only rely on the capacity of the human calculator, equipped with pencil and paper and with such aids as the slide rule, tables of logarithms, and other convenient tables.
  The advent of electronic computer removes the restriction of the human operator.
- The estimates defined by nonlinear equations can be established as a matter of routine by the appropriate iterative algorithms.

Examples:

- **Example 5.** Suppose $\theta = 0$ or 1 ($\Theta = \{0, 1\}$) and $p(x, \theta)$ is given by the following table.

$p(x, \theta)$	$x = 0$	$x = 2$
$\theta = 1$	0	1
$\theta = 2$	0.1	0.9

  Suppose that we observe two observations 2 and 2.
  How do we get them?
  Abstraction:

- X: a discrete random variable with pmf $p(x, \theta)$
- 2 and 2 are realizations of X_1 and X_2.
- What is the pmf of (X_1, X_2)?

Then

$$L(0, (2,2)) = 1 \qquad L(1, (2,2)) = (0.9)^2$$

and $\hat{\theta}((2,2)) = 0$.

- **Example 6.** If $x_1, \ldots, x_n$ are i.i.d. according to the Poisson distribution $\mathcal{P}(\lambda)$, the likelihood is

$$L(\lambda, \mathbf{x}) = \lambda^{\sum_i x_i} e^{-n\lambda} / \prod_i x_i!.$$

This is maximized by

$$\hat{\lambda} = \sum_i x_i / n$$

which is therefore the MLE of λ.
In this example, $\Theta = (0, \infty)$ and $k = 1$.
Use $rpois(20, 3)$ to generate 20 observations from $\mathcal{P}(3)$. They are $2, 3, 3, 5, 6, 3, 0, 5, 3, 2, 2, 2, 2, 2, 4, 1, 3, 2$
Then $\hat{\lambda} = 3.1$.
Why do we need to introduce $\mathbf{X}$?

- **Example 7.** Let $X_1, \ldots, X_n$ be i.i.d. according to the uniform distribution $U(0, \theta)$, so that the likelihood is

$$L(\theta, \mathbf{x}) = \begin{cases} 1/\theta^n & \text{if } 0 \leq x_i \leq \theta \text{ for all } i \\ 0 & \text{otherwise.} \end{cases}$$

  We can no longer differentiate $L(\theta, \mathbf{x})$ to get the MLE.
  By direct maximization, the MLE is equal to $x_{(n)}$.

- **Example 8.** Consider n items whose times to failure $X_1, \ldots, X_n$ form a sample from an $\mathcal{E}(\theta)$ distribution. (i.e., $p(x) = \theta \exp(-\theta x)$ for $x > 0$)
  Suppose the items are inspected only at discrete times $1, 2, \ldots, k$ so that we really observe $Y_1, \ldots, Y_n$ where, for $j = 1, \ldots, n$,

$$\begin{aligned} Y_j &= \ell \quad \text{if } \ell - 1 < X_j \leq \ell,\ \ell = 1, \ldots, k \\ &= k + 1 \quad \text{if } X_j > k. \end{aligned}$$

  Suppose $n = 20$, $k = 5$, and $\theta = 3$. x_is are 5.19, 0.06, 2.37, 4.38, 4.98, 13.02, 0.34, 7.26, 0.67, 1.96, 3.82, 0.27, 1.83, 3.48, 3.03, 1.90, 6.42, 7.49, 5.67, 6.27 and y_is are 6, 1, 3, 5, 5, 6, 1, 6, 1, 2, 4, 1, 2, 4, 4, 2, 6, 6, 6, 6.

Let N_i = number of indices j such that $Y_j = i$, $i = 1, \ldots, k+1$. Then the multinomial vector $\mathbf{N} = (N_1, \ldots, N_{k+1})$ is sufficient for θ and the likelihood function of $\mathbf{N}$ is

$$L(\theta, n_1, \ldots, n_{k+1}) = \frac{n!}{n_1! \cdots n_{k+1}!} \prod_{j=1}^{k+1} p_j^{n_j}(\theta),$$

where $p_j(\theta) = \exp(-[j-1]\theta) - \exp(-j\theta)$ for $1 \le j \le k$ and $p_{k+1}(\theta) = \exp(-k\theta)$. Question: How do we solve this problem?

Limitations on MLE

- It is a constant theme of the history of the method that the use of ML techniques is not always accompanied by a clear appreciation of their limitations.

- **Example 9.** (Neyman-Scott (1948) problem)
  In this example, the MLE is not even consistent.
  Refer to J. Neyman and E.L. Scott. Consistent estimate based on partially consistent observations. *Econometrica* **16** 1-32 (1948).
  Estimation of a Common Variance:

Let $X_{\alpha j}$ $(j = 1, \ldots, r)$ be independently distributed according to $N(\theta_\alpha, \sigma^2)$, $\alpha = 1, \ldots, n$. The MLEs are

$$\hat{\theta}_\alpha = X_{\alpha\cdot}, \quad \hat{\sigma}^2 = \frac{1}{rn} \sum_{\alpha=1}^{n} \sum_{j=1}^{r} (X_{\alpha j} - X_{\alpha\cdot})^2.$$

Furthermore, these are the unique solutions of the likelihood equations.

However, in the present case, the MLE of σ^2 is not even consistent.

To see this, note that the statistics

$$S_\alpha^2 = \sum_{j=1}^{r} (X_{\alpha j} - X_{\alpha\cdot})^2$$

are identically independently distributed with expectation

$$E(S_\alpha^2) = (r-1)\sigma^2$$

so that $\Sigma S_\alpha^2/n \to (r-1)\sigma^2$ and hence

$$\hat{\sigma}^2 \to \frac{r-1}{r}\sigma^2 \text{ in probability.}$$

- **Example 10.** (Non-existence of MLE)
If $Y_1, \ldots, Y_n$ are i.i.d. according to the Poisson distribution $P(\lambda)$. Suppose for each i we observe only when Y_i is 0 or positive and

let

$$X_i = \begin{cases} 0 & \text{if } Y_i = 0 \\ 1 & \text{if } Y_i > 0. \end{cases}$$

Then

$P(X_i = 0) = \exp(-\lambda)$, $P(X_i = 1) = 1 - \exp(-\lambda)$,

and the likelihood is

$L(\lambda) = [1 - \exp(-\lambda)]^{\Sigma x_i} \exp(-\lambda \, \Sigma[1 - x_i])$.

This is maximized by

$$\hat{\lambda} = -\log(1 - \bar{x}),$$

provided $\Sigma(1 - x_i) > 0$.
When all the x's are $= 1$, the likelihood becomes

$$L(\lambda) = [1 - \exp(-\lambda)]^n,$$

which is an increasing function of λ. In this case, the likelihood does not take on its maximum for any finite λ and the MLE does not exist. (Does it make sense?)
Discussions:

– For any fixed n, the probability $P(X_1 = \cdots = X_n = 1) = (1 - \exp(-\lambda))^n$ tends to 1 as $\lambda \to \infty$. Thus there will exist values of λ for which the probability is close to 1 that the MLE is undefined.

- For any fixed λ, the probability $P(X_1 = \cdots = X_n = 1) = (1 - \exp(-\lambda))^n$ tends to 0 as $n \to \infty$.

Iterative Procedures

In applications MLE's typically do not have analytic forms and some numerical methods have to be used to compute MLE's.

It is usually possible to assume that MLE emerges as a solution of the *likelihood equations*. Namely,

$$\frac{\partial}{\partial \theta_i} \log p(\mathbf{x}, \boldsymbol{\theta}) = 0, \quad i = 1, \cdots, k.$$

Symbolically, the equations we have to solve may be written

$$D_{\boldsymbol{\theta}} \ell(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{0},$$

where $\ell(\mathbf{x}, \boldsymbol{\theta}) = \log p(\mathbf{x}, \boldsymbol{\theta})$ and $D_{\boldsymbol{\theta}}$ is the vector differential operator whose ith component is $\partial/\partial\theta_i$.

A commonly used numerical method is the Newton-Raphson iteration method.

- Solve the likelihood equation $L^{(1)}(\boldsymbol{\theta}, \mathbf{x}) = \mathbf{0}$ iteratively.

- Replace $L^{(1)}(\boldsymbol{\theta}, \mathbf{x})$ by the linear terms of its Taylor expansion about a starting value $\hat{\boldsymbol{\theta}}^{(0)}$.
- Replace the likelihood equation with the equation

  $$L^{(1)}(\hat{\boldsymbol{\theta}}^{(0)}, \mathbf{x}) + L^{(2)}(\hat{\boldsymbol{\theta}}^{(0)}, \mathbf{x})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{(0)}) = \mathbf{0}.$$

  The solution for $\boldsymbol{\theta}$ is

  $$\hat{\boldsymbol{\theta}}^{(1)} = \hat{\boldsymbol{\theta}}^{(0)} - [L^{(2)}(\hat{\boldsymbol{\theta}}^{(0)}, \mathbf{x})]^{-1} L^{(1)}(\hat{\boldsymbol{\theta}}^{(0)}, \mathbf{x}),$$

  as a first approximation to the solution of the likelihood equation.
- Iterative the above procedure by replacing $\hat{\boldsymbol{\theta}}^{(0)}$ by $\hat{\boldsymbol{\theta}}^{(1)}$ and so on.
- In general,

  $$\hat{\boldsymbol{\theta}}^{(t+1)} = \hat{\boldsymbol{\theta}}^{(t)} - \left.\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(t)}} \left[\left.\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(t)}} \right]^{-1}.$$

  – The laborious aspect of this iterative procedure is the inversion of the matrix $\partial^2 L(\boldsymbol{\theta}^{(t)})/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T$ at the tth stage.

  – If our initial approximation $\hat{\boldsymbol{\theta}}^{(0)}$ is good, then $\partial^2 L(\boldsymbol{\theta}^{(0)})/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T$ will be near $\partial^2 L(\boldsymbol{\theta}^{(t)})/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T$ in non-pathological conditions.

We can often use the former matrix *at each stage* of the procedure.

  - It often happens that terms awkward to calculate appear in $\partial^2 L(\boldsymbol{\theta}^{(t)})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$ but not in its expected value.
    Sometimes, we replace $\partial^2 L(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$ by its expected value $E[\partial^2 L(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T]$, where the expectation is taken under $P_{\boldsymbol{\theta}}$. This method is known as the Fisher-scoring method.
    In most instances, $E[\partial^2 L(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T]$ is simply the negative information matrix discussed in the second topic.

- Issues on implementation:
  - Specification of the starting point: To ensure a sequence $\hat{\boldsymbol{\theta}}^{(t)}$ which converges to $\hat{\boldsymbol{\theta}}$, it requires that $\hat{\boldsymbol{\theta}}^{(t)}$ is sufficiently close to the root $\hat{\boldsymbol{\theta}}$.
  - Take any estimator which satisfies $\sqrt{n}(\hat{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\theta})$ is bounded in probability.
  - Specification of the stopping rule:
- **Example 11.** Probit Analysis

– Suppose the probability $\pi(s)$ that an individual responds to the level s of a stimulus can be expressed in the form

$$\pi(s) = \Phi\left(\frac{s-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(x-\mu)/\sigma} e^{-z^2/2} dz.$$

– The level s_i of the stimulus is applied to n_i individuals $(i = 1, \ldots, r)$ and the numbers m_i $(i = 1, \ldots, r)$ of responses at the different levels are observed.

– Determine MLE of μ and σ.

– $\ell(\mathbf{x}, (\mu, \sigma)) = constant + \Sigma_i \{m_i \log \pi(s_i) + (n_i - m_i) \log(1 - \pi(s_i))\}$ and the likelihood equations are

$$\begin{aligned} \sum_i \frac{m_i - n_i\pi_i}{\pi_i(1-\pi_i)} \frac{\partial \pi(s_i)}{\partial \mu} &= 0, \\ \sum_i \frac{m_i - n_i\pi_i}{\pi_i(1-\pi_i)} \frac{\partial \pi(s_i)}{\partial \sigma} &= 0. \end{aligned}$$

– Obtain initial approximations μ_0 and σ_0 to their solution.

– Suppose the $\pi(s_i)$s are known.
The plot of the points $(s_i, \Phi^{-1}(\pi(s_i)))$

would lie on the straight line

$$\Phi^{-1}(\pi) = \frac{s-\mu}{\sigma}.$$

Since m_i/n_i is an estimate of $\pi(s_i)$, we can fit a straight line to this set of points to yield estimates of μ and σ.

– The Hessian matrix is a rather complicated expression.
If we use Fisher-scoring method, it is given by

$$\begin{bmatrix} \Sigma_i \frac{-n_i}{\pi_i(1-\pi_i)} \left(\frac{\partial \pi(s_i)}{\partial \mu}\right)^2 & \Sigma_i \frac{-n_i}{\pi_i(1-\pi_i)} \frac{\partial \pi(s_i)}{\partial \mu} \frac{\partial \pi(s_i)}{\partial \sigma} \\ \Sigma_i \frac{-n_i}{\pi_i(1-\pi_i)} \frac{\partial \pi(s_i)}{\partial \mu} \frac{\partial \pi(s_i)}{\partial \sigma} & \Sigma_i \frac{-n_i}{\pi_i(1-\pi_i)} \left(\frac{\partial \pi(s_i)}{\partial \sigma}\right)^2 \end{bmatrix}.$$

Method of Moments

- It is the oldest method of deriving point estimators.
  Proposed by Karl Pearson (1894).

- Consider a parametric problem where $X_1, \ldots, X_n$ are i.i.d. random variables from $P_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \Theta \subset R^k$.
  Suppose that $m_1(\boldsymbol{\theta}), \ldots, m_k(\boldsymbol{\theta})$ are the first k moments of the population we are sampling from.

$$m_j(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(X_1^j).$$

- Define the jth sample moment $\hat{m}_j$ by

$$\hat{m}_j = \frac{1}{n} \sum_{i=1}^{n} X_i^j = E_{F_n}(X^j).$$

- Suppose we want to estimate $q(\boldsymbol{\theta})$ which can be expressed as

$$q(\boldsymbol{\theta}) = g(m_1(\boldsymbol{\theta}), \ldots, m_k(\boldsymbol{\theta})),$$

  where g is a continuous function.

- The method of moments estimate of $q(\boldsymbol{\theta})$ is

$$T(\mathbf{X}) = g(\hat{m}_1, \ldots, \hat{m}_k).$$

- Basic ideas:
  - Law of Large Numbers:

$$\hat{m}_j \xrightarrow{P} m_j(\boldsymbol{\theta})$$

  - Continuity:

Example 12. Consider the estimation of μ and σ^2 if $X_1, \ldots, X_n$ are a random sample from a population with mean μ and variance σ^2.

Example 13.

- Normal Mixtures. Consider an industrial setting with a production process in control, so that the outcome follows a known distribution, which we shall take to be the standard normal distribution.

  However, it is suspected that the production process has become contaminated, with the contaminating portion following some other unknown normal distribution $N(\eta, \tau^2)$.

  A sample $x_1, \ldots, x_n$ of the output is drawn. The X's are therefore assumed to be i.i.d. according to the distribution

$$pN(0,1) + (1-p)N(\eta, \tau^2).$$

Apply the method of moments, we have

$$\begin{aligned} m_1(p,\eta,\tau) = E(X_i) &= (1-p)\eta, \\ m_2(p,\eta,\tau) = E(X_i^2) &= p + (1-p)(\eta^2+\tau^2), \\ m_3(p,\eta,\tau) = E(X_i^3) &= (1-p)\eta(\eta^2+3\tau^2) \end{aligned}$$

and thus obtain the estimating equations

$$\begin{aligned} (1-p)\eta &= \bar{x}, \\ p + (1-p)(\eta^2+\tau^2) &= n^{-1}\sum_i x_i^2, \\ (1-p)\eta(\eta^2+3\tau^2) &= n^{-1}\sum_i x_i^3. \end{aligned}$$

- Do you know how to express (p, η, τ) as functions of m_1, m_2 and m_3?
- In general, how can we know whether the above task is possible?
- Implicit Function Theorem in advanced calculus.

- For the above example, suppose $\tau = 1$. The resulting estimators of η and $1-p$ are

$$\hat{\eta} = \frac{n^{-1}\Sigma_i X_i^2 - 1}{\bar{X}}, \qquad \text{and} \quad 1-\hat{p} = \frac{\bar{X}^2}{n^{-1}\Sigma_i X_i^2 - 1}.$$

The Frequency Substitution Principle

- Suppose we observe n multinomial trials in which the values $v_1, \ldots, v_k$ of the population being sampled are known, but their respective probabilities $p_1, \ldots, p_k$ are completely unknown.
- Let N_i denote the number of indices j such that $X_j = v_i$. Then $(N_1, \ldots, N_k)$ has a multinomial distribution with parameter $(n, p_1, \ldots, p_k)$. Here $\Sigma_i N_i = n$ and n is any natural number while $(p_1, \ldots, p_k)$ is any vector in
$$\{(p_1, \ldots, p_k) : p_i \geq 0, \sum_i p_i = 1\}.$$
- If $(N_1, \ldots, N_k)$ has a $\mathcal{M}(n, p_1, \ldots, p_k)$,
$$p(n_1, \ldots, n_k) = \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k},$$
$E(N_i) = np_i$, $Var(N_i) = np_i(1 - p_i)$, and $Cov(N_i, N_j) = -np_ip_j$ for $i \neq j$.
- The intuitive estimate of p_i is N_i/n, the proportion of sample values equal to v_i.
- Suppose we want to estimate a continuous function $q(p_1, \ldots, p_k)$.

The frequency substitution principle will give the estimate by replacing the unknown population frequencies $p_1, \dots, p_k$ by the observable sample frequencies $N_1/n, \dots, N_k/n$. That is

$$T(N_1, \dots, N_k) = q(N_1/n, \dots, N_k/n).$$

- Basic ideas:
  - Law of Large Numbers:

    $$N_j/n \xrightarrow{P} p_j$$

  - Continuity:
    A function f is said to be continuous at x_0 if $f(x_0+)$ and $f(x_0-)$ exist and if

    $$f(x_0+) = f(x_0-) = f(x_0).$$

    Refer to any advanced calculus book for details.

Example 14. Estimation in 2×2 tables

- Consider n independent trials, the outcome of each classified according to two criteria, as A or $\bar{A}$, and as B or $\bar{B}$.
  For example, a series of operations is being

classified according to the gender of the patient and the success or failure of the treatment.

- The results can be displayed in a 2×2 table as show below.

	B	$\bar{B}$	
A	n_{AB}	$n_{A\bar{B}}$	n_A
$\bar{A}$	$n_{\bar{A}B}$	$n_{\bar{A}\bar{B}}$	$n_{\bar{A}}$
	n_B	$n_{\bar{B}}$	n

  where n_{AB} is the number of cases having both attributes A and B, and so on.

- The joint distribution of the four cell entries is then multinomial with parameters $(n, p_{AB}, p_{\bar{A}B}, p_{A\bar{B}}, p_{\bar{A}\bar{B}})$.

- A standard measure of the degree of association of the attributes A and B is the cross-product ratio (also called odds ratio)

$$\rho = \frac{p_{AB} p_{\bar{A}\bar{B}}}{p_{\bar{A}B} p_{A\bar{B}}}.$$

  - Use the fact that $p_{AB} = p_A p_{B|A}$ where p_A and $p_{B|A}$ denote the probability of A and the conditional probability of B given A, respectively. It leads to

$$\rho = \frac{p_{B|A} p_{\bar{B}|A}}{p_{B|\bar{A}} p_{\bar{B}|\bar{A}}}.$$

- Think of A as the maternal age is no more than 20, $\bar{A}$ as the maternal age is greater than 20, B as the birthweight is no more than $2,500 gms$, and $\bar{B}$ as the birthweight is greater than $2,500 gms$. The odds ratio can be used to associate the risk of underweight baby to the maternal age.
- The attributes A and B are said to be positively associated if
$$p_{B|A} > p_{B|\bar{A}} \text{ and } p_{\bar{B}|\bar{A}} > p_{\bar{B}|A},$$
and these conditions imply that $\rho > 1$.
- In the case of negative dependence, the above inequalities are reversed.
- Independence of A and B is characterized by equality instead of inequality and hence by $\rho = 1$.

- The odds ratio ρ is estimated by replacing the cell probabilities $p_{AB}, \ldots$ by the corresponding frequencies $n_{AB}/n, \ldots$.

The Method of Least Squares

- It became widely used early in the nineteenth century by Gauss for estimation in problems of astronomical measurement.
- Suppose that water is being pumped through a container to which an amount of dye has been added.
  Every few seconds the concentration of dye is measured in the water leaving the container.
  It is expected that the concentration of dye will decrease linearly over time.
  Since the measuring equipment maynot be perfectly accurate, it may not be possible to interpret the measurements exactly, and the mixing may not behave exactly as predicted.
  The determine the rate at which the concentration decreases, the experimenter would have to approximate the data by a straight line, a line that *best* approximated the data in some sense. A common approach is to employ the method of least squares.
- The model above is called a **linear model**

because it is a linear combination of the model functions 1 and x.
x refers to the concentration of dye.
The model can be written as

$$Y_x = \theta_1 + \theta_2 x + \epsilon_x,$$

where Y_x is often called the response observed at x, (θ_1, θ_2) is a 2-vector of unknown parameters, x is an explanatory variable (or covariate), and ϵ_x is random error.
Our data is (x, Y_x) and ϵ_x cannot be observed.
x can be random or nonrandom.

- *Nonlinear* models are also used. A common example is an exponential model such as

$$Y_t = \theta_1 \exp(\theta_2 t) + \epsilon_t.$$

Here the model is a nonlinear function of the parameter β.

- In either case, we can write the observations $(\mathbf{x}_i, y_i)'s$ in the form,

$$Y_i = g_i(\theta_1, \ldots, \theta_k) + \epsilon_i, \quad 1 \leq i \leq n.$$

where

- The g_i are known functions and the real numbers $\theta_1, \ldots, \theta_k$ are unknown parameters of interest.
- The parameters $(\theta_1, \ldots, \theta_k)$ can vary freely over a set Ω contained in R^k.
- The ϵ_i satisfy the following restriction:

$$\begin{aligned} E(\epsilon_i) &= 0, \quad 1 \le i \le n, \\ Var(\epsilon_i) &= \sigma^2, \quad 1 \le i \le n, \\ Cov(\epsilon_i, \epsilon_j) &= 0, \quad 1 \le i < j \le n. \end{aligned}$$

- $E(Y_i) = g_i(\theta_1, \ldots, \theta_k)$ with unknown $\theta_1, \ldots, \theta_k$.

Example 15.

- Suppose that we want to find out how increasing the amount x of a certain chemical or fertilizer in the soil increases the amount y of that chemical in the plants grown in that soil.
  - Nine samples of soil were treated with different amounts x of phosphorus.
  - Y is the amount of phosphorus found in corn plants grown for 38 days in the different samples of soil.

- (x_i, y_i) are $(1, 64)$, $(4, 71)$, $(5, 54)$, $(9, 81)$, $(11, 76)$, $(13, 93)$, $(23, 77)$, $(23, 95)$, $(28, 109)$.

- Assume the relationship between x and y can be approximated well by a random model $y_i = \theta_1 + \theta_2 x_i + \epsilon_i$.

- A least squares estimator of (θ_1, θ_2) is defined to be the minimizer of

$$Q(\theta_1, \theta_2) = \sum_{i=1}^{9} (y_i - \theta_1 - \theta_2 x_i)^2.$$

We then run into an optimization problem. Note that

  - $Q(\theta_1, \theta_2)$ is a quadratic function of (θ_1, θ_2).
  - There is no restriction on the ranger of (θ_1, θ_2). (i.e., $(\theta_1, \theta_2) \in R^2$ which falls in an open set.)

    It follows from vector calculus that the least squares estimate $(\hat{\theta}_1, \hat{\theta}_2)$ must satisfy the equations

$$\frac{\partial}{\partial \theta_j} Q(\theta_1, \theta_2) = 0, \quad j = 1, 2.$$

    If the constraint is imposed, we may need to use the method of Lagrange multiplier to find the minimizer.

– Differentiation leads to the following normal equations

$$\begin{aligned} \sum_i (y_i - \theta_1 - \theta_2 x_i) &= 0 \\ \sum_i x_i (y_i - \theta_1 - \theta_2 x_i) &= 0. \end{aligned}$$

The sample regression line is $61.58 + 1.42x$.

- If some of $\{\epsilon_1, \cdots, \epsilon_n\}$ have more chance of being small than others it might seem more sensible to estimate θ_1 and θ_2 by minimizing some weighted sum of squares
$$\sum_{i=1}^{n} w_i (y_i - \theta_1 - \theta_2 x_i)^2,$$
the ws being weights which are larger for those is for which ϵ_i is liable to be small and small for ϵ_i liable to be large.

Optimization and Least Squares

- The word **optimization** denotes either the minimization or maximization of a function.

- Consider a real-valued function h with domain D in R^k. The function h is said to have a *local maximum* at point $\theta^* \in D$ if there exists a real number $\delta > 0$ such that $h(\theta) \leq h(\theta^*)$ for all $\theta \in D$ satisfying $\|\theta - \theta^*\| \leq \delta$.

Define a *local minimum* in a similar way, but in the sense that inequality $h(\theta) \leq h(\theta^*)$ is reversed.

If the inequality $h(\theta) \leq h(\theta^*)$ is replaced by a strict inequality

$$h(\theta) < h(\theta^*), \quad \theta \in D, \theta \neq \theta^*,$$

we have a strict local maximum; and if the sense of the inequality $h(\theta) < h(\theta^*)$ is reversed, we have a strict local minimum.

- We say that the function h has a *global (absolute) maximum* (strict global maximum) at θ^* if $h(\theta) \leq h(\theta^*)$, $[h(\theta) < h(\theta^*)]$ holds for every $\theta \in D$.

  Thus a function may have many local maxima, each with a different value of $h(\theta)$, say, $h(\theta_j^0)$, $j = 1, \ldots, \ell$.

  The global maximum can always be chosen from among these local maxima by comparing their values and choosing one such that

$$h(\theta^*) \geq h(\theta_j^0), \quad j = 1, \ldots, \ell,$$

where $\theta^* \in \{\theta_j^0, j = 1, \ldots, \ell\}$.
It is clear that every global maximum (minimum) is also a local maximum (minimum); however, the converse of this statement is, in general, not true.
Only when $h(\theta)$ is a convex function in R^k and $D \subset R^k$ is a convex set is every local extremum of h at $\theta \in D$ also a global extremum of h over D.

- Minimization of a one-dimensional function $h(\theta)$, without any restrictions on θ, by Newton's method:
  - Assume that $h(\theta)$ has at least two continuous derivatives and that it is bounded below.
  - Approximate $h(\theta)$ by a quadratic function that we can minimize, and use the minimizer of the simpler function as the new estimate of the minimizer of $h(\theta)$. The process is then repeated from this new point.
  - To form a quadratic approximation, let $\theta^{(t)}$ be the current estimate of the solu-

tion θ^*, and consider a Taylor series expansion of h about the point $\theta^{(t)}$:

$$h(\theta^{(t)}+s) = h(\theta^{(t)})+sh^{'}(\theta^{(t)})+\frac{1}{2}s^2h^{''}(\theta^{(t)})+\cdots.$$

The original minimization problem can be approximated using a Taylor series expansion

$$\begin{aligned} h(\theta^*) &= \min_\theta h(\theta) = \min_s h(\theta^{(t)}+s) \\ &= \min_s \left[h(\theta^{(t)}) + sh^{'}(\theta^{(t)}) + \frac{1}{2}s^2h^{''}(\theta^{(t)}) + \cdots\right] \\ &\approx \min_s \left[h(\theta^{(t)}) + sh^{'}(\theta^{(t)}) + \frac{1}{2}s^2h^{''}(\theta^{(t)})\right]. \end{aligned}$$

– To minimize the quadratic, take the derivative with respect to s and set it equal to zero giving

$$s = -\frac{h^{'}(\theta^{(t)})}{h^{''}(\theta^{(t)})}.$$

Since s is an approximation to the step that would take us from $\theta^{(t)}$ to the solution θ^* of the original problem, and the algorithm is defined by the formula

$$\theta^{(t+1)} = \theta^{(t)} - \frac{h^{'}(\theta^{(t)})}{h^{''}(\theta^{(t)})}.$$

- Optimization in many dimensions with linear regression
  - Consider Example 15 in which $Q(\theta_1, \theta_2)$ can be written as $(\theta_1, \theta_2) \begin{pmatrix} 9 & \Sigma_i x_i \\ \Sigma_i x_i & \Sigma_i x_i^2 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} - 2(\theta_1, \theta_2) \begin{pmatrix} \Sigma_i y_i \\ \Sigma_i x_i y_i \end{pmatrix} + \sum_i y_i^2.$
  - How do we differentiate a quadratic form $\boldsymbol{\theta}^T \mathbf{A} \boldsymbol{\theta}$?
    Here $\mathbf{A}$ is a $k \times k$ square matrix and symmetric.
    Result:
    $$\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\theta}^T \mathbf{A} \boldsymbol{\theta} = 2\mathbf{A}\boldsymbol{\theta}.$$
  - How do we differentiate $\boldsymbol{\theta}^T \mathbf{b}$?
    Here $\mathbf{b}$ is a $k \times 1$ column vector.
    Result:
    $$\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\theta}^T \mathbf{b} = \mathbf{b}.$$
  - Matrix formulation of the linear model:
    $$\mathbf{y} = \mathcal{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}.$$
    Here $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\mathcal{X} = (x_{ij})_{n \times k}$, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$. Observe that
    $(\mathbf{y} - \mathcal{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathcal{X}\boldsymbol{\theta}) = \boldsymbol{\theta}^T \mathcal{X}^T \mathcal{X} \boldsymbol{\theta} - 2\boldsymbol{\theta}^T \mathcal{X}^T \mathbf{y} + \mathbf{y}^T \mathbf{y}.$

Differentiation leads to the normal equations

$$2\mathcal{X}^T\mathcal{X}\boldsymbol{\theta} - 2\mathcal{X}^T\mathbf{y} = 0,$$

Any solution of the above is an LSE of $\boldsymbol{\theta}$. If $\mathcal{X}$ is of full rank, in which case $(\mathcal{X}^T\mathcal{X})^{-1}$ exists, then there is a unique LSE which is

$$\hat{\boldsymbol{\theta}} = (\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^T\mathbf{y}.$$

- In R, the function *solve* inverts matrices and solve systems of linear equations; $solve(A)$ inverts A and $solve(A, b)$ solves $A\% * \%x = b$.

  If the system is over-determined, the least-squares fit is found, but matrices of less than full rank give an error.
- Consider the simple linear regression. It turns out that

$$\mathcal{X}^T\mathcal{X} = \begin{pmatrix} n & \Sigma_i\, x_i \\ \Sigma_i\, x_i & \Sigma_i\, x_i^2 \end{pmatrix}$$

  The matrix is invertible if and only if some x_i's are different.

- Optimization in Many Dimensions: Newton's Method

- Newton's method (also called the Newton-Raphson method) is a widely used and often-studied method for minimization.
- The method requires use of both the gradient vector and the Hessian matrix in computations; hence it places more burden on the user to supply derivatives of the objective function than does the steepest descent method learned in calculus (the gradient vector defines the direction of maximum local increase).
- Write the Taylor series in matrix/vector form. In two dimensions, the second-order Taylor series approximation is

$$\begin{aligned} h(\theta_1 + s_1, \theta_2 + s_2) &\approx h(\theta_1, \theta_2) + s_1 D^{(1,0)} h(\theta_1, \theta_2) + s_2 D \\ &+ \frac{1}{2} \Big[s_1^2 D^{(2,0)} h(\theta_1, \theta_2) + 2 s_1 s_2 D^{(1,1)} h \\ &\quad + s_2^2 D^{(0,2)} h(\theta_1, \theta_2) \Big] . \end{aligned}$$

- Let $\bigtriangledown^2 h$ be the constant matrix of second partial derivatives of h at θ_j-the so-called **Hessian matrix**:

$$\bigtriangledown^2 h_{ij} = \frac{\partial^2 h(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}.$$

– If the notations for the gradient and Hessian matrix are used, we can write down Taylor series in many dimensions which takes the form

$$Q(\boldsymbol{\theta}+\mathbf{p}) = h(\boldsymbol{\theta})+\mathbf{p}^T \bigtriangledown h(\boldsymbol{\theta})+\frac{1}{2}\mathbf{p}^T \bigtriangledown^2 h(\boldsymbol{\theta})\mathbf{p}.$$

– When $\boldsymbol{\theta}^{(t)}$ is close to $\boldsymbol{\theta}^*$ we can expect that the above quadratic function will approximate $h(\boldsymbol{\theta})$.
To obtain the step $\mathbf{p}$, we now minimize this quadratic as a function of $\mathbf{p}$ by forming its gradient with respect to $\mathbf{p}$

$$\begin{aligned} \bigtriangledown_{\mathbf{p}} Q(\mathbf{p}) &= \bigtriangledown_{\mathbf{p}} \left(\mathbf{p}^T \bigtriangledown h(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{p}^T \bigtriangledown^2 h(\boldsymbol{\theta})\mathbf{p}\right) \\ &= \bigtriangledown h(\boldsymbol{\theta}) + \bigtriangledown^2 h(\boldsymbol{\theta})\mathbf{p} \end{aligned}$$

and setting it equal to zero

$$\bigtriangledown^2 h(\boldsymbol{\theta})\mathbf{p} = - \bigtriangledown h(\boldsymbol{\theta}).$$

This is a set of n linear equations in the k unknowns $\mathbf{p} = (p_1, \ldots, p_k)^T$.
These linear equations are called the **Newton equations**.
If $\bigtriangledown^2 h(\boldsymbol{\theta})$ is positive definite, this sug-

gests the general iterative scheme

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \mathbf{p} = \boldsymbol{\theta}^{(t)} - [\bigtriangledown^2 h(\boldsymbol{\theta}^{(t)})]^{-1} \bigtriangledown h(\boldsymbol{\theta}).$$

- When $h(\boldsymbol{\theta})$ is closely approximated by $Q(\mathbf{p})$ in the neighborhood of $\boldsymbol{\theta}^*$, convergence will normally be at a quadratic rate if the Hessian is positive definite at each step.
- One problem with Newton's method is that the Hessian may not be positive definite at each iteration.

  Thus the method requires modification to insure that the resultant method is acceptable but still retains the desirable characteristics of Newton's method.

Recall the nonlinear regression. If a least-squares approach were used, the following optimization problem would be obtained

$$\min_{\theta_0, \theta_1} \sum_{i=1}^{n} [Y_i - \theta_0 \exp(\theta_1 T_i)]^2.$$

This is called a **nonlinear least-squares problem**. No analytic solution can be found. More details will be given when we discuss MLE later on.

Prediction

- Suppose we have a random vector (or variable) $\mathbf{X}$ with $E\mathbf{X}^T\mathbf{X} < \infty$ and a random variable Y.
  One may wish to predict the value of Y based on an observed value of $\mathbf{X}$.
  Let $g(\mathbf{X})$ be the predictor with $E[g(\mathbf{X})]^2 < \infty$.

- As a motivated example, a stock holder wants to predict the value of his holdings at some time in the future (Y) on the basis of his past experience with the market and his portfolio ($\mathbf{X}$).

- Suppose we use a linear function of $\mathbf{X}$ (instead of nonlinear function) to predict of Y. What is the best linear predictor under mean squared error?

  - Suppose that $E(X^2)$ and $E(Y^2)$ are finite and X and Y are not constant. Then the unique best zero intercept linear predictor is obtained by taking

    $$a = a_0 = E(XY)/E(X^2),$$

while the unique best linear predictor is $a_1X + b_1$ where

$$a_1 = Cov(X, Y)/Var(X), \quad b_1 = E(Y) - a_1E(X).$$

- If we don't find any good predictor, how do we predict Y?
  We may use $E(Y)$. (Note that $E(Y)$ is the constant which minimizes $E(Y - c)^2$.)
- How do we predict Y if we use the least absolute criterion $E|Y - c|$?
- Up to now, we have three possible predictor of Y.
  They are sample mean, linear predictor, and conditional mean (smoother).
- In general, we have to find θ_0 and $\boldsymbol{\theta}$ to minimize

$$E(Y - \theta_0 - \boldsymbol{\theta}^T\mathbf{X})^2.$$

A simple algebra leads to

$$\begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_k \end{pmatrix} = \begin{pmatrix} 1 & E(X_1) & \cdots & E(X_k) \\ E(X_1) & E(X_1^2) & \cdots & E(X_1X_k) \\ \vdots & \vdots & \vdots & \vdots \\ E(X_k) & E(X_kX_1) & \cdots & E(X_k^2) \end{pmatrix}^{-1} \begin{pmatrix} E(Y) \\ E(X_1Y \\ \vdots \\ E(X_kY \end{pmatrix}$$

Later on, we will compare this to its sample version.

- Do we need condition to ensure the above matrix is invertible?

  Do you remember the concept of positive definite?

  When $E(X) = 0$, what can we say about X?

Basic ideas on the method of least squares:

- Substitution principle:

  Suppose we view the linear regression as a best linear predictor problem.

  The linear regression is the one that minimizes the following estimated mean squared error:

$$n^{-1} \sum_{i=1}^{n} (y_i - \theta_0 - \boldsymbol{\theta}^T \mathbf{x}_i)^2.$$

- Is the above estimated mean squared error a good estimate of $E(Y - \theta_0 - \boldsymbol{\theta}^T \mathbf{X})^2$?

- We can use standard argument involving the linear combination of random variable by computing its mean and variance of least squares solution.