Statistical models and estimation[1]

STA431 Spring 2017

[1] See last slide for copyright information.

Overview

1 Models

2 MOM

3 MLE

4 Invariance

5 Consistency

Statistical model

Most good statistical analyses are based on a *model* for the data.

A *statistical model* is a set of assertions that partly specify the probability distribution of the observable data. The specification may be direct or indirect.

- Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with expected value μ and variance σ^2.
- For $i = 1, \ldots, n$, let $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i$, where

  $\beta_0, \ldots, \beta_k$ are unknown constants.
  x_{ij} are known constants.
  $\epsilon_1, \ldots, \epsilon_n$ are independent $N(0, \sigma^2)$ random variables, not observable.
  σ^2 is an unknown constant.
  $Y_1, \ldots, Y_n$ are observable random variables.

A model is not the same thing as the *truth*.

Statistical models leave something unknown

Otherwise they are probability models

- The unknown part of the model for the data is called the *parameter*.
- Usually, parameters are (vectors of) numbers.
- Usually denoted by θ or $\boldsymbol{\theta}$ or other Greek letters.
- Parameters are unknown constants.

Parameter Space

The *parameter space* is the set of values that can be taken on by the parameter.

- Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with expected value μ and variance σ^2. The parameter space is $\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$.
- For $i = 1, \ldots, n$, let $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i$, where

  $\beta_0, \ldots, \beta_k$ are unknown constants.
  x_{ij} are known constants.
  $\epsilon_1, \ldots, \epsilon_n$ are independent $N(0, \sigma^2)$ random variables.
  σ^2 is an unknown constant.
  $Y_1, \ldots, Y_n$ are observable random variables.

  The parameter space is $\Theta = \{(\beta_0, \ldots, \beta_k, \sigma^2) : -\infty < \beta_j < \infty, \sigma^2 > 0\}$.

Parameters need not be numbers

Let $X_1, \ldots, X_n$ be a random sample from a continuous distribution with unknown distribution function $F(x)$.

- The parameter is the unknown distribution function $F(x)$.
- The parameter space is a space of distribution functions.
- We may be interested only in a *function* of the parameter, like
$$\mu = \int_{-\infty}^{\infty} x f(x) \, dx$$

The rest of $F(x)$ is just a nuisance parameter.

General statement of a statistical model

D is for Data

$$D \sim P_\theta, \quad \theta \in \Theta$$

- Both D and θ could be vectors
- For example,
  - $D = \mathbf{Y}_1, \ldots \mathbf{Y}_n$ independent multivariate normal.
  - $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
  - P_θ is the joint distribution function of $\mathbf{Y}_1, \ldots \mathbf{Y}_n$, with joint density

$$f(\mathbf{y}_1, \ldots \mathbf{y}_n) = \prod_{i=1}^{n} f(\mathbf{y}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Estimation

For the model $D \sim P_\theta, \quad \theta \in \Theta$

- We don't know θ.
- We never know θ.
- All we can do is guess.
- Estimate θ (or a function of θ) based on the observable data.
- T is an *estimator* of θ (or a function of θ): $T = T(D)$

For example,

- $D = X_1, \ldots, X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$ $\quad T = (\overline{X}, S^2)$.
- For an ordinary multiple regression model, $T = (\widehat{\boldsymbol{\beta}}, MSE)$

T is a *statistic*, a random variable (vector) that can be computed from the data without knowing the values of any unknown parameters.

Parameter estimation

For the model $D \sim P_\theta, \quad \theta \in \Theta$

- Estimate θ with $T = T(D)$.
- How do we get a recipe for T? Guess?
- It's good to be systematic. Lots of methods are available.
- We will consider two: Method of moments and maximum likelihood.

Moments

Based on a random sample like $(X_1, Y_1), \ldots, (X_n, Y_n)$

- Moments are quantities like $E\{X_i\}$, $E\{X_i^2\}$, $E\{X_iY_i\}$, $E\{W_iX_i^2Y_i^3\}$, etc.
- *Central* moments are moments of *centered* random variables:

$$
\begin{array}{l}
E\{(X_i - \mu_x)^2\} \\
E\{(X_i - \mu_x)(Y_i - \mu_y)\} \\
E\{(X_i - \mu_x)^2(Y_i - \mu_y)^3(Z_i - \mu_z)^2\}
\end{array}
$$

- These are all *population* moments.

Population moments and sample moments

Population moment	Sample moment
$E\{X_i\}$	$\frac{1}{n}\sum_{i=1}^n X_i$
$E\{X_i^2\}$	$\frac{1}{n}\sum_{i=1}^n X_i^2$
$E\{X_iY_i\}$	$\frac{1}{n}\sum_{i=1}^n X_iY_i$
$E\{(X_i-\mu_x)^2\}$	$\frac{1}{n}\sum_{i=1}^n(X_i-\overline{X}_n)^2$
$E\{(X_i-\mu_x)(Y_i-\mu_y)\}$	$\frac{1}{n}\sum_{i=1}^n(X_i-\overline{X}_n)(Y_i-\overline{Y}_n)$
$E\{(X_i-\mu_x)(Y_i-\mu_y)^2\}$	$\frac{1}{n}\sum_{i=1}^n(X_i-\overline{X}_n)(Y_i-\overline{Y}_n)^2$

Estimation by the Method of Moments (MOM)

For the model $D \sim P_\theta, \quad \theta \in \Theta$

- Population moments are a function of θ.
- Find θ as a function of the population moments.
- Estimate θ with that function of the *sample* moments.

Symbolically,

- Let m denote a vector of population moments.
- $\widehat{m}$ is the corresponding vector of sample moments.
- Find $m = g(\theta)$
- Solve for θ, obtaining $\theta = g^{-1}(m)$.
- Let $\widehat{\theta} = g^{-1}(\widehat{m})$.

It doesn't matter if you solve first or put hats on first.

Example: $X_1, \ldots, X_n \overset{i.i.d}{\sim} U(0, \theta)$

$f(x) = \frac{1}{\theta}$ for $0 < x < \theta$

First find the moment (expected value).

$$\begin{aligned} E(X_i) &= \int_0^\theta x \frac{1}{\theta}\, dx \\ &= \frac{1}{\theta} \int_0^\theta x\, dx \\ &= \frac{1}{\theta}\ \frac{x^2}{2}\bigg|_0^\theta = \frac{1}{2\theta}(\theta^2 - 0) \\ &= \frac{\theta}{2} \end{aligned}$$

So $m = \frac{\theta}{2} \Leftrightarrow \theta = 2m$, and $\widehat{\theta} = 2\overline{X}$.

Small numerical example

Let $X_1, \ldots, X_n$ be a random sample from a uniform distribution on $(0, \theta)$. Estimate θ by the Method of Moments for the following data. Your answer is a number. Show some work.

```
4.09 0.13 0.84 3.83 2.13 4.67 4.61 0.40 4.19 0.71
```

$\overline{X} = 2.56$ so $\widehat{\theta} = 2\overline{X} = 2 * 2.56 = 5.12$.

Method of moments estimators are not unique

What moments you use are up to you.

$$E(X_i^2) = \frac{1}{\theta}\int_0^\theta x^2\,dx = \frac{\theta^2}{3}$$

So set $m = \frac{\theta^2}{3} \Leftrightarrow \theta = \sqrt{3m}$, and

$$\widehat{\theta} = \sqrt{\frac{3}{n}\sum_{i=1}^{n} X_i^2}$$

Compared to $2\overline{X}$.

Compare $\widehat{\theta}_1 = 2\overline{X}$ and $\widehat{\theta}_2 = \sqrt{\frac{3}{n}\sum_{i=1}^{n} X_i^2}$

For the numerical example

x	4.09	0.13	0.84	3.83	2.13	4.67	4.61	0.40	4.19
x^2	16.7281	0.0169	0.7056	14.6689	4.5369	21.8089	21.2521	0.16	17.5561

$$\widehat{\theta}_1 = 5.12 \qquad \widehat{\theta}_2 = 5.42$$

Expressions for lower order moments tend to be simpler, and are preferable if only for that reason.

Method of Moments estimator for normal

Let $X_1, \ldots, X_n \stackrel{i.i.d}{\sim} N(\mu, \sigma^2)$

From the moment-generating function or a textbook, $E(X_i) = \mu$ and $E(X_i^2) = \sigma^2 + \mu^2$. Solving for the parameters,

$$
\begin{aligned}
\mu &= E(X_i) \\
\sigma^2 &= E(X_i^2) - (E(X_i))^2
\end{aligned}
$$

so

$$
\begin{aligned}
\widehat{\mu} &= \overline{X} \\
\widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}^2 \\
&= \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2
\end{aligned}
$$

A regression example

Independently for $i = 1, \ldots, n$,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \text{ where}$$

- $E(X_i) = \mu_x$, $Var(X_i) = \sigma^2_x$
- $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2_\epsilon$
- X_i and ϵ_i are independent.
- The distributions of X_i and ϵ_i are unknown.
- What's the parameter?

- The parameter is $(\beta_0, \beta_1, F_\epsilon(\epsilon), F_x(x)\,)$.
- We want to estimate β_0 and β_1, a *function* of the parameter.

Calculate some moments

$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

$$\begin{aligned}
&E(X_i) = \mu_x \\
&Var(X_i) = \sigma_x^2 \\
&E(Y_i) = \beta_0 + \beta_1 \mu_x \\
&Cov(X_i, Y_i) = \beta_1 \sigma_x^2
\end{aligned}$$

Use the centering rule to get the last one:

$$\begin{aligned}
Cov(X_i, Y_i) &= E(\overset{c}{X}_i \overset{c}{Y}_i) \\
&= E\{\overset{c}{X}_i \, (\beta_1 \, \overset{c}{X}_i + \epsilon_i)\} \\
&= E\{\beta_1 \, \overset{c}{X}{}_i^2 + \overset{c}{X}_i \, \epsilon_i)\} \\
&= \beta_1 E\{\overset{c}{X}{}_i^2\} + E\{\overset{c}{X}_i\} E\{\epsilon_i\} \\
&= \beta_1 \sigma_x^2
\end{aligned}$$

Solve for β_0 and β_1

Have $E(X_i) = \mu_x$, $Var(X_i) = \sigma_x^2$, $E(Y_i) = \beta_0 + \beta_1\mu_x$, $Cov(X_i, Y_i) = \beta_1\sigma_x^2$

Putting hats on first, solve

$$\begin{aligned} \overline{Y} &= \widehat{\beta}_0 + \widehat{\beta}_1\overline{X} \\ \widehat{\sigma}_{xy} &= \widehat{\beta}_1\widehat{\sigma}_x^2 \end{aligned}$$

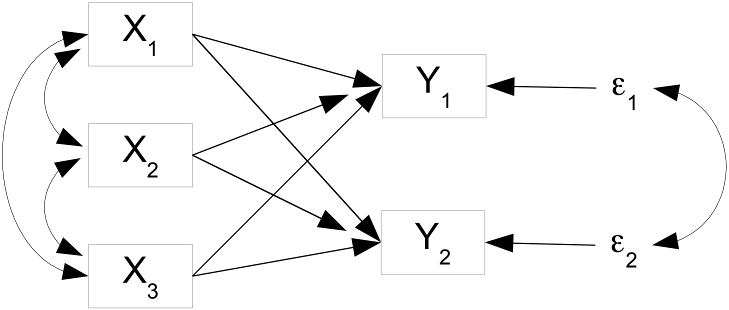
$\Rightarrow$

$$\begin{aligned} \widehat{\beta}_1 &= \frac{\widehat{\sigma}_{xy}}{\widehat{\sigma}_x^2} = \frac{\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^n (X_i - \overline{X}_n)^2} \text{ and} \\ \widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1\overline{X} \end{aligned}$$

These happen to be the same as the least-squares estimates.

Multivariate multiple regression

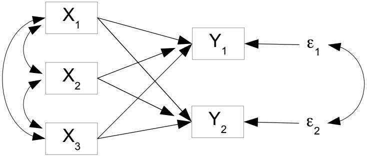
Multivariate means more than one response variable



We will obtain method of moments estimation for this.

One regression equation for each response variable

Give the equations in scalar form.



$$
\begin{aligned}
Y_{i,1} &= \beta_{1,0} + \beta_{1,1}X_{i,1} + \beta_{1,2}X_{i,2} + \beta_{1,3}X_{i,3} + \epsilon_{i,1} \\
Y_{i,2} &= \beta_{2,0} + \beta_{2,1}X_{i,1} + \beta_{2,2}X_{i,2} + \beta_{2,3}X_{i,3} + \epsilon_{i,2}
\end{aligned}
$$

$\mathbf{Y}_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{X}_i + \boldsymbol{\epsilon}_i$

In scalar form,

$$\begin{aligned} Y_{i,1} &= \beta_{1,0} + \beta_{1,1} X_{i,1} + \beta_{1,2} X_{i,2} + \beta_{1,3} X_{i,3} + \epsilon_{i,1} \\ Y_{i,2} &= \beta_{2,0} + \beta_{2,1} X_{i,1} + \beta_{2,2} X_{i,2} + \beta_{2,3} X_{i,3} + \epsilon_{i,2} \end{aligned}$$

In matrix form,

$$\begin{array}{cccccccc} \mathbf{Y}_i & = & \boldsymbol{\beta}_0 & + & \boldsymbol{\beta}_1 & \mathbf{X}_i & + & \boldsymbol{\epsilon}_i \\ \begin{pmatrix} Y_{i,1} \\ Y_{i,2} \end{pmatrix} & = & \begin{pmatrix} \beta_{1,0} \\ \beta_{2,0} \end{pmatrix} & + & \begin{pmatrix} \beta_{1,1} & \beta_{1,2} & \beta_{1,3} \\ \beta_{2,1} & \beta_{2,2} & \beta_{2,3} \end{pmatrix} & \begin{pmatrix} X_{i,1} \\ X_{i,2} \\ X_{i,3} \end{pmatrix} & + & \begin{pmatrix} \epsilon_{i,1} \\ \epsilon_{i,2} \end{pmatrix} \end{array}$$

Note different order from $Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i$

Statement of the model

Independently for $i = 1, \ldots, n$,

$$\mathbf{Y}_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{X}_i + \boldsymbol{\epsilon}_i, \text{ where}$$

- $\mathbf{Y}_i$ is an $q \times 1$ random vector of observable response variables, so the regression is multivariate; there are q response variables.
- $\mathbf{X}_i$ is a $p \times 1$ observable random vector; there are p explanatory variables. $E(\mathbf{X}_i) = \boldsymbol{\mu}_x$ and $cov(\mathbf{X}_i) = \boldsymbol{\Phi}_{p \times p}$. The vector $\boldsymbol{\mu}_x$ and the matrix $\boldsymbol{\Phi}$ are unknown.
- $\boldsymbol{\beta}_0$ is a $q \times 1$ vector of unknown constants.
- $\boldsymbol{\beta}_1$ is a $q \times p$ matrix of unknown constants. These are the regression coefficients, with one row for each response variable and one column for each explanatory variable.
- $\boldsymbol{\epsilon}_i$ is a $q \times 1$ unobservable random vector with expected value zero and unknown variance-covariance matrix $cov(\boldsymbol{\epsilon}_i) = \boldsymbol{\Psi}_{q \times q}$.
- $\boldsymbol{\epsilon}_i$ is independent of $\mathbf{X}_i$.

A Method of Moments estimate of $\boldsymbol{\beta}_1$

$\mathbf{Y}_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{X}_i + \boldsymbol{\epsilon}_i$

Denote the $p \times q$ matrix of (population) covariances between $\mathbf{X}_i$ and $\mathbf{Y}_i$ by

$$
\begin{aligned}
\boldsymbol{\Sigma}_{xy} &= cov(\mathbf{X}_i, \mathbf{Y}_i) \\
&= E\{\overset{c}{\mathbf{X}}_i \overset{c}{\mathbf{Y}}_i^\top\} \\
&= E\{\overset{c}{\mathbf{X}}_i \, (\boldsymbol{\beta}_1 \, \overset{c}{\mathbf{X}}_i \, + \boldsymbol{\epsilon}_i)^\top\} \\
&= E\{\overset{c}{\mathbf{X}}_i \; (\overset{c}{\mathbf{X}}_i^\top \boldsymbol{\beta}_1^\top + \boldsymbol{\epsilon}_i^\top)\} \\
&= E\{\overset{c}{\mathbf{X}}_i \overset{c}{\mathbf{X}}_i^\top \boldsymbol{\beta}_1^\top + \overset{c}{\mathbf{X}}_i \; \boldsymbol{\epsilon}_i^\top\} \\
&= E\{\overset{c}{\mathbf{X}}_i \overset{c}{\mathbf{X}}_i^\top\} \boldsymbol{\beta}_1^\top + E\{\overset{c}{\mathbf{X}}_i \; \boldsymbol{\epsilon}_i^\top\} \\
&= cov(\mathbf{X}_i) \boldsymbol{\beta}_1^\top + cov(\mathbf{X}_i, \boldsymbol{\epsilon}_i) \\
&= \boldsymbol{\Phi} \boldsymbol{\beta}_1^\top + \mathbf{0} \\
&= \boldsymbol{\Phi} \boldsymbol{\beta}_1^\top
\end{aligned}
$$

Solve for $\boldsymbol{\beta}_1$

In terms of moments of the observable data

$$
\begin{array}{rrcr}
 & \boldsymbol{\Phi}\boldsymbol{\beta}_1^\top & = & \boldsymbol{\Sigma}_{xy} \\
\Rightarrow & \boldsymbol{\Phi}^{-1}\boldsymbol{\Phi}\boldsymbol{\beta}_1^\top & = & \boldsymbol{\Phi}^{-1}\boldsymbol{\Sigma}_{xy} \\
\Rightarrow & \boldsymbol{\beta}_1^\top & = & \boldsymbol{\Phi}^{-1}\boldsymbol{\Sigma}_{xy} \\
\Rightarrow & \boldsymbol{\beta}_1 & = & \boldsymbol{\Sigma}_{xy}^\top(\boldsymbol{\Phi}^{-1})^\top \\
 & & = & \boldsymbol{\Sigma}_{yx}\boldsymbol{\Phi}^{-1} \\
 & & = & \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_x^{-1},
\end{array}
$$

Where $\boldsymbol{\Phi} = cov(\mathbf{X}_i)$ is written $\boldsymbol{\Sigma}_x$.

MOM estimate of $\boldsymbol{\beta}_1$ based on $\boldsymbol{\beta}_1 = \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_x^{-1}$

Just put hats on.

$$\widehat{\boldsymbol{\beta}}_1 = \widehat{\boldsymbol{\Sigma}}_{yx}\widehat{\boldsymbol{\Sigma}}_x^{-1},$$

where

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}}_{yx} &= \frac{1}{n}\sum_{i=1}^n (\mathbf{Y}_i - \overline{\mathbf{Y}})(\mathbf{X}_i - \overline{\mathbf{X}})^\top \\
\widehat{\boldsymbol{\Sigma}}_{x} &= \frac{1}{n}\sum_{i=1}^n (\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^\top
\end{aligned}$$

Method of Moments is Least Squares in this case

$$\widehat{\boldsymbol{\beta}}_1 = \widehat{\boldsymbol{\Sigma}}_{yx}\widehat{\boldsymbol{\Sigma}}_x^{-1}$$

- This is $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$
- Transposed
- With both x and y variables centered by subtracting off the sample means.

Maximum likelihood estimation

A great idea from R. A. Fisher (1890-1962)

- Given a model and a set of observed data, how should we estimate θ?
- Find the value of θ that makes the data we observed have the highest probability.
- If the model is continuous, maximize the probability of observing data in a little region surrounding the observed data vector.
- In either case, let $f(\mathbf{d};\theta)$ denote the joint probability density function or probability mass function evaluated at the observed data vector.
- Maximize $L(\theta) = f(\mathbf{d};\theta)$ over all $\theta \in \Theta$.
- $L(\theta)$ is called the *likelihood function*.

Maximum likelihood estimation for independent random sampling

$$D_1, \ldots, D_n \overset{i.i.d.}{\sim} P_\theta,\ \theta \in \Theta.$$

$$L(\theta) = \prod_{i=1}^{n} f(d_i;\theta),$$

where $f(d_i;\theta)$ is the density or probability mass function evaluated at d_i.

- Find the value of θ for which $L(\theta)$ is maximum.
- Or equivalently, maximize $\ell(\theta) = \ln L(\theta)$.
- The elementary approach:
  - Take derivatives,
  - Set derivatives to zero,
  - Solve for θ,
  - Put a hat on the answer.

Example: Coffee taste test

A fast food chain is considering a change in the blend of coffee beans they use to make their coffee. To determine whether their customers prefer the new blend, the company plans to select a random sample of $n = 100$ coffee-drinking customers and ask them to taste coffee made with the new blend and with the old blend, in cups marked "A" and "B." Half the time the new blend will be in cup A, and half the time it will be in cup B. Management wants to know if there is a difference in preference for the two blends.

Statistical model for the taste test example

Letting θ denote the probability that a consumer will choose the new blend, treat the data $Y_1, \ldots, Y_n$ as a random sample from a Bernoulli distribution. That is, independently for $i = 1, \ldots, n$,

$$f(y_i; \theta) = \theta^{y_i}(1-\theta)^{1-y_i}$$

for $y_i = 0$ or $y_i = 1$, and zero otherwise.

Find the MLE of θ

Show your work

Maximize the log likelihood.

$$
\begin{aligned}
\frac{\partial}{\partial\theta}\ln L(\theta) &= \frac{\partial}{\partial\theta}\ln\left(\prod_{i=1}^{n} f(y_i;\theta)\right) \\
&= \frac{\partial}{\partial\theta}\ln\left(\prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}\right) \\
&= \frac{\partial}{\partial\theta}\ln\left(\theta^{\sum_{i=1}^{n} y_i}(1-\theta)^{n-\sum_{i=1}^{n} y_i}\right) \\
&= \frac{\partial}{\partial\theta}\left((\sum_{i=1}^{n} y_i)\ln\theta + (n-\sum_{i=1}^{n} y_i)\ln(1-\theta)\right) \\
&= \frac{\sum_{i=1}^{n} y_i}{\theta} - \frac{n-\sum_{i=1}^{n} y_i}{1-\theta}
\end{aligned}
$$

Setting the derivative to zero,

$$
\begin{aligned}
\frac{\sum_{i=1}^n y_i}{\theta} = \frac{n - \sum_{i=1}^n y_i}{1-\theta} \quad & \Rightarrow \quad (1-\theta)\sum_{i=1}^n y_i = \theta(n - \sum_{i=1}^n y_i) \\
& \Rightarrow \quad \sum_{i=1}^n y_i - \theta \sum_{i=1}^n y_i = n\theta - \theta \sum_{i=1}^n y_i \\
& \Rightarrow \quad \sum_{i=1}^n y_i = n\theta \\
& \Rightarrow \quad \theta = \frac{\sum_{i=1}^n y_i}{n}
\end{aligned}
$$

So it looks like the MLE is the sample proportion. Carrying out the second derivative test to be sure,

Second derivative test

$$
\begin{aligned}
\frac{\partial^2 \ln \ell}{\partial \theta^2} &= \frac{\partial}{\partial \theta}\left(\frac{\sum_{i=1}^n y_i}{\theta} - \frac{n - \sum_{i=1}^n y_i}{1-\theta} \right) \\
&= \frac{-\sum_{i=1}^n y_i}{\theta^2} - - - \frac{n - \sum_{i=1}^n y_i}{(1-\theta)^2} \\
&= -n \left(\frac{1-\overline{y}}{(1-\theta)^2} + \frac{\overline{y}}{\theta^2} \right) < 0
\end{aligned}
$$

Concave down, maximum, verifying $\widehat{\theta} = \overline{y}$.

Numerical estimate

Suppose 60 of the 100 consumers prefer the new blend. Give a point estimate the parameter θ. Your answer is a number.

```
> ybar = 60/100; ybar
[1] 0.6
```

Maximum likelihood for the univariate normal

Let $X_1, \ldots, X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$.

$$\begin{aligned}
\ell(\theta) &= \ln \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}} \\
&= \ln\left(\sigma^{-n}(2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}\right) \\
&= -n\ln\sigma - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2
\end{aligned}$$

Differentiate with respect to the parameters

$\ell(\theta) = -n \ln \sigma - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2$

$$\begin{aligned}
\frac{\partial \ell}{\partial \mu} &= -\frac{1}{2\sigma^2}\sum_{i=1}^n 2(x_i - \mu)(-1) \overset{set}{=} 0 \\
&\Rightarrow \mu = \overline{x}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \ell}{\partial \sigma} &= -\frac{n}{\sigma} - \frac{1}{2}\sum_{i=1}^n (x_i - \mu)^2(-2\sigma^{-3}) \\
&= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^n (x_i - \mu)^2 \overset{set}{=} 0 \\
&\Rightarrow \sigma^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \mu)^2
\end{aligned}$$

Substituting

Setting derivaties to zero, we have obtained

$$\mu = \overline{x} \text{ and } \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2, \text{ so}$$

$$\begin{aligned} \widehat{\mu} &= \overline{X} \\ \widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \end{aligned}$$

Gamma Example

Let $X_1, \ldots, X_n$ be a random sample from a Gamma distribution with parameters $\alpha > 0$ and $\beta > 0$

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

$$\Theta = \{(\alpha, \beta) : \alpha > 0, \beta > 0\}$$

Log Likelihood

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

$$\begin{aligned}
\ell(\alpha,\beta) &= \ln\prod_{i=1}^{n}\frac{1}{\beta^\alpha\Gamma(\alpha)}e^{-x_i/\beta}x_i^{\alpha-1} \\
&= \ln\left(\beta^{-n\alpha}\,\Gamma(\alpha)^{-n}\exp(-\frac{1}{\beta}\sum_{i=1}^{n}x_i)\left(\prod_{i=1}^{n}x_i\right)^{\alpha-1}\right) \\
&= -n\alpha\ln\beta - n\ln\Gamma(\alpha) - \frac{1}{\beta}\sum_{i=1}^{n}x_i + (\alpha-1)\sum_{i=1}^{n}\ln x_i
\end{aligned}$$

Differentiate with respect to the parameters

$\ell(\theta) = -n\alpha \ln \beta - n \ln \Gamma(\alpha) - \frac{1}{\beta} \sum_{i=1}^n x_i + (\alpha - 1) \sum_{i=1}^n \ln x_i$

$$
\begin{aligned}
\frac{\partial \ell}{\partial \beta} &\overset{set}{=} 0 \quad \Rightarrow \quad \alpha\beta = \overline{x} \\
\frac{\partial \ell}{\partial \alpha} &= -n \ln \beta - n \frac{\partial}{\partial \alpha} \ln \Gamma(\alpha) + \sum_{i=1}^n \ln x_i \\
&= \sum_{i=1}^n \ln x_i - n \ln \beta - n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \overset{set}{=} 0
\end{aligned}
$$

Solve for α

$$\sum_{i=1}^{n} \ln x_i - n \ln \beta - n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = 0$$

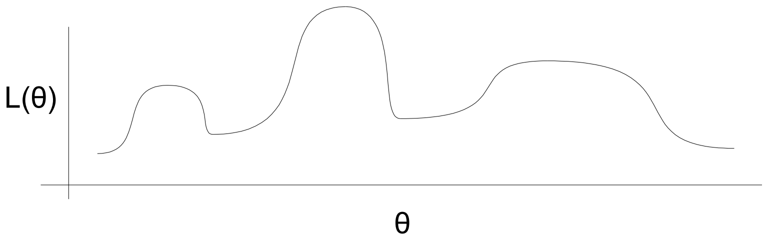
where

$$\Gamma(\alpha) = \int_0^\infty e^{-t} t^{\alpha-1}\, dt.$$

Nobody can do it.

Maximize the likelihood numerically with software

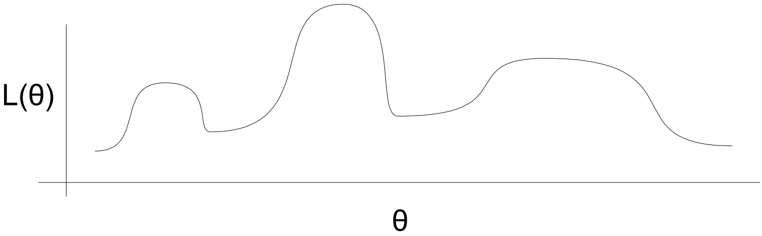
Usually this is in high dimension



- It’s like trying to find the top of a mountain by walking uphill blindfolded.
- You might stop at a local maximum.
- The starting place is very important.
- The final answer is a number (or vector of numbers).
- There is no explicit formula for the MLE.

There is a lot of useful theory

Even without an explicit formula for the MLE



- MLE is asymptotically normal.
- Variance of the MLE is deeply related to the curvature of the log likelihood at the MLE.
- The more curvature, the smaller the variance.
- The variance of the MLE can be estimated from the curvature (using the Fisher Information).
- Basis of tests and confidence intervals.

Comparing MOM and MLE

- Sometimes they are identical, sometimes not.
- If the model is right they are usually close for large samples.
- Both are asymptotically normal.
- Estimates of the variance are easy to obtain for both.
- Small variance of an estimator is good.
- As $n \to \infty$, nothing can beat the MLE.
- Except that the MLE depends on a very specific distribution.
- And sometimes the dependence matters.
- In such cases, MOM is preferable.

The Invariance principle of maximum likelihood estimation

Also applies to Method of Moments estimation

- The Invariance Principle of maximum likelihood estimation says that *the MLE of a function is that function of the MLE, provided the function is one-to-one.*
- An example comes first, followed by formal details.

Example

Of the invariance principle

Let $D_1, \ldots, D_n$ be a random sample from a Bernoulli distribution (1=Yes, 0=No) with parameter $\theta, 0 < \theta < 1$. The parameter space is $\Theta = (0,1)$, and the likelihood function is

$$L(\theta) = \prod_{i=1}^{n} \theta^{d_i}(1-\theta)^{1-d_i} = \theta^{\sum_{i=1}^n d_i}(1-\theta)^{n-\sum_{i=1}^n d_i}.$$

Differentiating the log likelihood with respect to θ, setting the derivative to zero and solving yields the usual estimate $\widehat{\theta} = \overline{d}$, the sample proportion.

Re-parameterize

- Write the model in terms of the *odds* of $D_i = 1$, a re-parameterization that is often useful in categorical data analysis.
- Denote the odds by θ'.
- The definition of odds is

$$\theta' = \frac{\theta}{1-\theta} = g(\theta).$$

- As θ ranges from zero to one, θ' ranges from zero to infinity.
- So there is a new parameter space: $\theta' \in \Theta' = (0, \infty)$.

Likelihood function in terms of $\theta' = \frac{\theta}{1-\theta}$

First solve for θ, obtaining $\theta = \frac{\theta'}{1+\theta'} = g^{-1}(\theta')$. The likelihood in terms of θ' is then

$$\begin{aligned}
L(g^{-1}(\theta')) &= \theta^{\sum_{i=1}^n d_i}(1-\theta)^{n-\sum_{i=1}^n d_i} \\
&= \left(\frac{\theta'}{1+\theta'}\right)^{\sum_{i=1}^n d_i}\left(1-\frac{\theta'}{1+\theta'}\right)^{n-\sum_{i=1}^n d_i} \\
&= \left(\frac{\theta'}{1+\theta'}\right)^{\sum_{i=1}^n d_i}\left(\frac{1+\theta'-\theta'}{1+\theta'}\right)^{n-\sum_{i=1}^n d_i} \\
&= \frac{\theta'^{\sum_{i=1}^n d_i}}{(1+\theta')^n}.
\end{aligned}$$

$L(g^{-1}(\theta')) = L'(\theta')) = \frac{\theta'^{\sum_{i=1}^n d_i}}{(1+\theta')^n}$

See how re-parameterization changes the likelihood function

- Could differentiate the log likelihood, set the derivative to zero, and solve for θ'.
- The point of the invariance principle is that this is unnecessary.
- The maximum likelihood estimator of $g(\theta) = \frac{\theta}{1-\theta}$ is $g(\widehat{\theta})$, so that

$$\widehat{\theta'} = \frac{\widehat{\theta}}{1-\widehat{\theta}} = \frac{\overline{d}}{1-\overline{d}} \; .$$

Theorem

See text for a proof. The one-to-one part is critical.

Let $g : \Theta \to \Theta'$ be a one-to-one re-parameterization, with the maximum likelihood estimate $\widehat{\theta}$ satisfying $L(\widehat{\theta}) > L(\theta)$ for all $\theta \in \Theta$ with $\theta \neq \widehat{\theta}$. Then $L'(g(\widehat{\theta})) > L'(\theta')$ for all $\theta' \in \Theta'$ with $\theta' \neq g(\widehat{\theta})$.

In other words

- The MLE of $g(\theta)$ is $g(\widehat{\theta})$.
- $\widehat{g(\theta)} = g(\widehat{\theta})$.
- The MLE of θ' is $g(\widehat{\theta})$.
- $\widehat{\theta}' = g(\widehat{\theta})$.

Re-parameterization in general

The parameters of common statistical models are written in a standard way, but other equivalent parameterizations are sometimes useful. Suppose $X_i \sim N(\mu, \sigma^2)$. Have

$$\widehat{\theta} = (\overline{X}, \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2)$$

- Write $X_i \sim N(\mu, \sigma)$.
  - $g(\theta) = (\theta_1, \sqrt{\theta_2})$
  - $\widehat{\theta}' = \left(\overline{X}, \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}\right)$
- Write $X_i \sim N(\mu, \tau)$, where $\tau = 1/\sigma^2$ is called the *precision*.
  - $g(\theta) = (\theta_1, 1/\theta_2)$
  - $\widehat{\theta}' = \left(\overline{X}, \frac{n}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right)$

Consistency

- The idea is large-sample accuracy.
- As $n \to \infty$, you get the truth.
- It's a kind of limit, but with probability involved.

The setting

- Let $T_1, T_2, \ldots$ be a sequence of random variables.
- Main application: T_n is an estimator of θ based on a sample of size n.
- Think $T_n = \overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$.
- Generalize to random vectors, soon.

Convergence in Probability

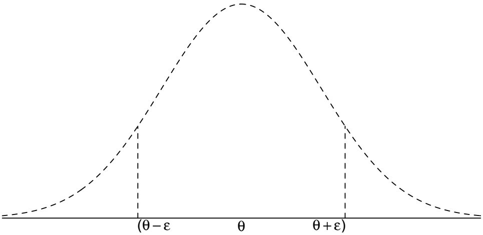
We say that T_n converges *in probability* to the constant θ, and write $T_n \xrightarrow{p} \theta$ if for all $\epsilon > 0$,

$$\lim_{n\to\infty} P\{|T_n - \theta| < \epsilon\} = 1$$

Convergence in probability to θ means no matter how small the interval around θ, for large enough n (that is, for all $n > N$) the probability of getting a value of T_n that near to θ (or nearer) is as close to one as you like.

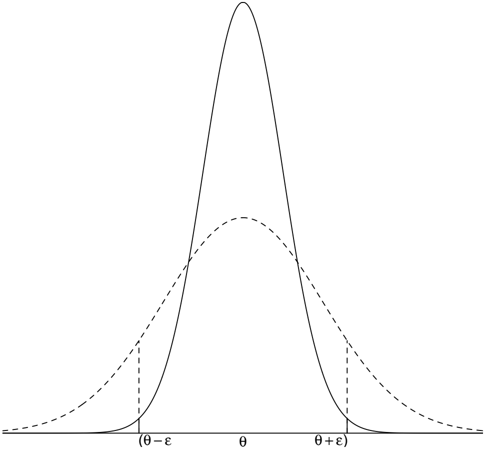
Picture it

$$\begin{aligned} P\{|T_n - t| < \epsilon\} &= P\{-\epsilon < T_n - \theta < \epsilon\} \\ &= P\{\theta - \epsilon < T_n < \theta + \epsilon\} \\ &. \end{aligned}$$



Picture it

$$\begin{aligned} P\{|T_n - t| < \epsilon\} &= P\{-\epsilon < T_n - \theta < \epsilon\} \\ &= P\{\theta - \epsilon < T_n < \theta + \epsilon\} \end{aligned}$$



Convergence in Probability for Random Vectors

Let $\mathbf{T}_1, \mathbf{T}_2, \ldots$ be a sequence of k-dimensional random vectors.

We say that $\mathbf{T}_n$ converges in probability to $\boldsymbol{\theta} \in \mathbb{R}^k$, and write $\mathbf{T}_n \stackrel{p}{\rightarrow} \boldsymbol{\theta}$ if for all $\epsilon > 0$,

$$\lim_{n \to \infty} P\{||\mathbf{T}_n - \boldsymbol{\theta}|| < \epsilon\} = 1,$$

where $||\mathbf{a} - \mathbf{b}||$ denotes Euclidian distance in $\mathbb{R}^k$.

Use theorems, not the definition

- In this class we will *not* use the definition of convergence in probability.
- We will use theorems instead.

The Law of Large Numbers

Let $X_1, X_2, \ldots$ be independent random variables from a distribution with expected value μ. The Law of Large Numbers says

$$\overline{X}_n \xrightarrow{p} \mu$$

The Change of Variables formula: Let $Y = g(X)$

$$E(Y) = \int_{-\infty}^{\infty} y \, f_{_Y}(y) \, dy = \int_{-\infty}^{\infty} g(x) \, f_{_X}(x) \, dx$$

Or, for discrete random variables

$$E(Y) = \sum_y y \, p_{_Y}(y) = \sum_x g(x) \, p_{_X}(x)$$

This is actually a big theorem, not a definition.

Applying the change of variables formula

To approximate $E[g(X)]$

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} g(X_i) &= \frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{p} E(Y) \\
&= E(g(X))
\end{aligned}
$$

So for example

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k \stackrel{p}{\rightarrow} E(X^k)$$

$$\frac{1}{n}\sum_{i=1}^{n} U_i^2 V_i W_i^3 \stackrel{p}{\rightarrow} E(U^2 V W^3)$$

- That is, sample moments converge in probability to population moments.
- Central sample moments converge to central population moments as well.

Two more Theorems

- The “stack” theorem and continuous mapping.
- Often used together.

The "Stack" Theorem

Because I don't know what to call it.

Let $\mathbf{X}_n \stackrel{p}{\rightarrow} \mathbf{x}$ and $\mathbf{Y}_n \stackrel{p}{\rightarrow} \mathbf{y}$. Then the partitioned random vector

$$\left(\begin{array}{c} \mathbf{X}_n \\ \mathbf{Y}_n \end{array} \right) \stackrel{p}{\rightarrow} \left(\begin{array}{c} \mathbf{x} \\ \mathbf{y} \end{array} \right)$$

Continuous mapping

One of the Slutsky lemmas

Let $\mathbf{T}_n \xrightarrow{p} \mathbf{t}$, and let the function $g(\mathbf{x})$ be continuous at $\mathbf{x} = \mathbf{t}$. Then

$$g(\mathbf{T}_n) \xrightarrow{p} g(\mathbf{t})$$

Note that the function g could be multidimensional, for example mapping $\mathbb{R}^5$ into $\mathbb{R}^2$.

Definition of Consistency

The random vector (of statistics) $\mathbf{T}_n$ is said to be a *consistent* estimator of the parameter vector $\boldsymbol{\theta}$ if

$$\mathbf{T}_n \xrightarrow{p} \boldsymbol{\theta}$$

for all $\boldsymbol{\theta} \in \Theta$.

Consistency of the Sample Variance

This answer gets full marks.

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$$

By LLN, $\overline{X}_n \xrightarrow{p} \mu$ and $\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{p} E(X^2) = \sigma^2 + \mu^2$.

By continuous mapping,

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 \xrightarrow{p} \sigma^2 + \mu^2 - \mu^2 = \sigma^2$$

Note the silent use of the Stack Theorem.

Method of Moments Estimators are Consistent

For most practical cases

Recall

- Let m denote a vector of population moments.
- $\widehat{m}$ is the corresponding vector of sample moments.
- Find $m = g(\theta)$
- Solve for θ, obtaining $\theta = g^{-1}(m)$.
- Let $\widehat{\theta}_n = g^{-1}(\widehat{m}_n)$.

If g is continuous, so is g^{-1}. Then by continous mapping, $\widehat{m} \xrightarrow{p} m \Rightarrow \widehat{\theta}_n = g^{-1}(\widehat{m}_n) \xrightarrow{p} g^{-1}(m) = \theta$.

Consistency is great but it's not enough.

- It's the least we can ask. Estimators that are *not* consistent are completely unacceptable for most purposes.
- Think of $a_n = 1/n$ as a sequence of degenerate random variables with $P\{a_n = 1/n\} = 1$.
- So, $a_n \stackrel{p}{\rightarrow} 0$.

$$T_n \stackrel{p}{\rightarrow} \theta \Rightarrow U_n = T_n + \frac{100,000,000}{n} \stackrel{p}{\rightarrow} \theta.$$

Copyright Information