Statistics Lecture

Claudio Ortelli [1]

[1]Finance Institute
Università della Svizzera italiana

Advanced Learning and Research Institute, 2011

Conditional Distribution and Expectation

Let A and B be two events and $P(B) \neq 0$. The conditional probability of the event A, given that event B is realized, is by definition

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

Let X be a random variable and define B to be the event that $X = x$. The conditional probability $P(A \mid X = x)$ of the event A is then

$$P(A \mid X = x) = \frac{P(A \text{ and } X = x)}{P(X = x)} = \frac{P(A \cap [X = x])}{P(X = x)} =$$

provided of course that $P(X = x) \neq 0$.

Conditional Distribution and Expectation

Definition

1) Conditional pmf. Let X and Y be discrete random variables with joint pmf $p(x,y)$. The conditional pmf of Y given X is

$$
\begin{aligned}
p_{Y|X}(y \mid x) &= P(Y = y \mid X = x) \\
&= \frac{P(Y = y, X = x)}{P(X = x)} = \frac{p(x,y)}{p_x(x)}
\end{aligned}
$$

if $p_X(x) \neq 0$ and 0 otherwise.

2) The conditional distribution function of a random Variable Y (not necessarily descrete) given a discrete random variable X is

$$
F_{Y|X}(y \mid x) = P(Y \leq y \mid X = x) = \frac{P(Y \leq y \text{ and } X = x)}{P(X = x)}
$$

for all y and all x such that $P(X = x) \neq 0$.

Example

- Server cluster with two servers labeled A and B.
- Incoming jobs are independently routed to A and B with probability p and $q = 1 - p$, respectively.
- The number X of arriving jobs per unit of time is Poisson distributed with intensity λ.
- Determine the number of jobs, Y, received by server A, per unit of time.

$$P_{Y|X}(k,n) = \begin{cases} P_{Y|X}(Y=k, X=n) = \begin{pmatrix} n \\ k \end{pmatrix} p^k q^{n-k}, & 0 \le k \le n. \\ 0 & \text{otherwise.} \end{cases}$$

Conditional Distribution and Expectation

Example

(Continued) Recall that $P(X=n) = e^{-\lambda}\lambda^n/n!$ so that

$$
\begin{aligned}
p_Y(k) &= \sum_{n=k}^{\infty} p_{Y|X}(k \mid n) p_X(n) \\
&= \sum_{n=k}^{\infty} \binom{n}{k} p^k q^{n-k} \frac{e^{-\lambda}\lambda^n}{n!} \\
&= \lambda^k p^k e^{-\lambda} \sum_{n=k}^{\infty} \binom{n}{k} \frac{1}{n!} q^{n-k} \lambda^{n-k} \\
&= \frac{(\lambda p)^k}{k!} e^{-\lambda} \sum_{n=k}^{\infty} \frac{(q\lambda)^{n-k}}{(n-k)!}
\end{aligned}
$$

so that finally $p_Y(k) = \frac{(\lambda p)^k}{k!} e^{-\lambda} e^{q\lambda} = \frac{(\lambda p)^k}{k!} e^{-\lambda p}$, i.e. Y is Poisson distributed with intensity λp.

If X is a continuous random variable then $P(X = x) = 0$ for all $x \in \mathbb{R}$ so that the previous definition $\frac{P(Y=y,X=x)}{P(X=x)}$ of conditional probability is not satisfactory.
However when X and Y are jointly continuous we can define the conditional pdf of Y given X:

Definition

Let X and Y be continuous r.v. with joint pdf $f(x, y)$. The conditional density $f_{Y|X}$ is

$$f_{Y|X}(y \mid x) = \begin{cases} \frac{f(x,y)}{f_X(x)}, & \text{if } 0 < f_X(x) < \infty, \\ 0 & \text{otherwise.} \end{cases}$$

Conditional Distribution and Expectation

From the definition of conditional density it follows that

$$f(x,y) = f_{Y|X}(y \mid x) f_X(x) = f_{X|Y}(x \mid y) f_Y(y),$$

and if X and Y are independent, then

$$f(x,y) = f_X(x) f_Y(y).$$

Furthermore,

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx = \int_{-\infty}^{\infty} f_{Y|X}(y \mid x) f_X(x)dx$$

which is the continuous analog of the thorem of total probability.

Conditional Distribution and Expectation

- The conditional pdf can be used to obtain the conditional probability:

$$P(a \leq Y \leq b \mid X = x) = \int_a^b f_{Y|X}(y \mid x)dy, \qquad a \leq b.$$

- The conditional distribution function is defined analogously

$$\begin{aligned} F_{Y|X}(y \mid x) &= P(Y \leq y \mid X = x) \\ &= \frac{\int_{-\infty}^{y} f(x,t)dt}{f_X(x)} \\ &= \int_{-\infty}^{y} f_{Y|X}(t \mid x)dt. \end{aligned}$$

Conditional Distribution and Expectation

Example

Consider a series system of two *independent* components with respective lifetime distributions $X \sim EXP(\lambda_1)$ and $Y \sim EXP(\lambda_2)$. We are interested in the probability of envent A that component 2 causes the system failure, i.e.

$$P(A) = P(X \geq Y).$$

The conditional pdf is $F_{X|Y}(t,t) = P(X \leq t \mid Y = t) = F_X(t)$ by the independence of X and Y. By the total prob. theorem (continuous version)

$$\begin{aligned} P(A) &= \int_0^{\infty} P(X \geq t, Y = t) f_Y(t) dt \\ &= \int_0^{\infty} [1 - F_X(t)] f_Y(t) dt = \frac{\lambda_2}{\lambda_1 + \lambda_2}. \end{aligned}$$

Conditional Distribution and Expectation

Exercise

Consider the three-dimensional vector $X = (X_1, X_2, X_3)$ *having the following joint density function*

$$f_X(x_1, x_2, x_3) = \begin{cases} 6x_1 x_2^2 x_3, & \text{if } 0 \le x_1 \le 1,\ 0 \le x_x \le 1,\ 0 \le x_3 \le \sqrt{2}. \\ 0, & \text{otherwise.} \end{cases}$$

1. *Compute the conditional density functions* $f_{X_1,X_2|X_3}(x_1, x_2 \mid x_3)$ *and* $f_{X_3|X_1}(x_3 \mid x_1)$.
2. *Verify if the three random variables* X_1, X_2, X_3 *are independent.*

Exercise

X_1 *and* X_2 *are independent r. v. with Poisson distribution, having respective parameters* α_1 *and* α_2. *Show that the conditional* pmf *of* X_1 *given* $X_1 + X_2$, $p_{X_1|X_1+X_2}(X_1 = x_1 \mid X_1 + X_2 = y)$, *is binomial. Determine its parameters.*

Exercise

Let the execution times X and Y of two independent parallel processes be uniformly distributed over $(0, t_X)$ and $(0, t_Y)$, respectively, with $t_X \leq t_Y$. Find the probability that the former process finishes execution before the later.

- Consider a file server whose workload may be divided into r distinct classes.
- For a job of class i ($1 \leq i \leq r$) the CPU time is exponentially distributed with parameter λ_i.
- Let Y denote the service time of a job and let X be the job class. Then

$$f_{Y|X}(y \mid i) = \lambda_i e^{-\lambda_i y}, \qquad y > 0.$$

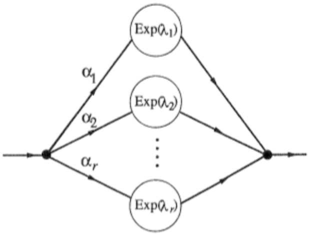
- Assume that the probability $p_X(i)$ that a randomly chosen job belongs to class i is equal to $\alpha_i > 0$. It follows $\sum_{i=1}^r \alpha_i = 1$.

The joint density of X and Y is then

$$f(i,y) = f_{Y|X}(y \mid i)p_X(i) = \alpha_i \lambda_i e^{-\lambda_i y}, \qquad y > 0.$$

The marginal density of Y is then

$$f_Y(y) = \sum_{i=1}^{r} f(i,y) = \sum_{i=1}^{r} \alpha_i \lambda_i e^{-\lambda_i y}, \qquad y > 0, \text{ i.e.}$$



Y has an $r-$stage hyperexponential distribution!

In general the conditional distribution of Y does not have to be exponential! Denoting $f_{Y|X}(y \mid i) = f_{Y_i}(y)$ and $F_{Y|X}(y \mid i) = F_i(y)$ then the unconditional pdf of Y is

$$f_Y(y) = \sum_{i=1}^{r} \alpha_i f_i(y)$$

and the unconditional CDF of Y is

$$F_Y(y) = \sum_{i=1}^{r} \alpha_i F_i(y).$$

Applying the definition of the mean and higher moments we obtain

$$\begin{aligned} E[Y] &= \sum_{i=1}^{r} \alpha_i E[Y_i], \\ E[Y^k] &= \sum_{i=1}^{r} \alpha_i E[Y_i^k]. \end{aligned}$$

Conditional Distribution and Expectation

- If X and Y are continuous random variables, we can for instance compute the conditional density $f_{Y|X}$.
- Since $f_{Y|X}$ has all properties of a density function of a continuous random variable, we can talk about its moments.
- Its mean (if exists) is called the conditional expectation of Y given $X = x$ and is denoted $E[Y \mid X = x]$ or $E[Y \mid x]$:

$$E[Y \mid x] \quad = \quad \begin{cases} \int_{-\infty}^{\infty} y f(y \mid x) dy, & \text{if } 0 < f(x) < \infty \\ 0 & \text{otherwise.} \end{cases}$$

- In case the random variables X and Y are discrete, $E[Y \mid x]$ is defined as

$$E[Y \mid X = x] = \sum_y yP(Y = y \mid X = x) = \sum_y y p_{Y|X}(y \mid x).$$

Conditional Distribution and Expectation

Similar arguments hold when X and Y are discrete. The conditional expectation is then defined as

$$E[Y \mid X = x] = \sum_y yP(Y = y \mid X = x) = \sum_y yp_{Y|X}(y \mid x).$$

Definition

The quantity

$$m(x) = E[Y \mid x]$$

considered as a function of x is known as the *regression function* of Y on X.

Definition

The conditional expectation of a function $\phi(Y)$ is defined as

$$E[\phi(Y) \mid X = x] = \begin{cases} \int_{-\infty}^{\infty} \phi(y) f_{Y|X}(y \mid x) dy, & \text{if } Y \text{ is continuous,} \\ \sum_i \phi(y_i) p_{Y|X}(y_i \mid x), & \text{if } Y \text{ is discrete.} \end{cases}$$

We may take expectation of the regression function to obtain the unconditional expectation of $\phi(Y)$

$$E[\phi(Y)] = \begin{cases} \sum_x E[\phi(Y) \mid X = x] p_X(x), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} E[\phi(Y) \mid X = x] f_X(x) dx, & \text{if } X \text{ is continuous.} \end{cases}$$

This last formula is known as the **theorem of total expectation**.

Limit Theorems

Theorem

(Chebyshev) Let X be a random variable with expected value μ and finite variance $\sigma^2 < \infty$. The, for all $t > 0$, the following inequality holds

$$P(| X - \mu | \geq t) \leq \frac{\sigma^2}{t^2}.$$

Definition

(Convergence in probability) Let $\{X_n\}_{n \in \mathbb{N}}$ be a sequence of random variables. We say that the sequence converges in probability to $c \in \mathbb{R}$, write $X_n \underset{p}{\rightarrow} c$ or $p\lim X_n = c$, if, for all $\varepsilon > 0$

$$\lim_{n \to \infty} P(| X_n - c | \geq \varepsilon) = 0.$$

Limit Theorems

Theorem

Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of random variables with common expectation μ and finite variance $\sigma_n^2 < \infty$. If $\lim_{n\to\infty} \sigma_n^2 = 0$, then

$$X_n \underset{p}{\rightarrow} \mu.$$

Proof.

Apply Chebyshev inequality. $\square$

Example

Let $\{X_n\}_{n\in\mathbb{N}}$ be an $i.i.d. \sim (\mu, \sigma^2)$ (independent and identically distributed) sequence of random variables. Define the sequence $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. From the linearity of the expectation and the properties of the variance we know that $\bar{X}_n \sim (\mu, \frac{\sigma^2}{n})$. The sequence $\{\bar{X}_n\}_{n\in\mathbb{N}}$ converges in probability to μ: $p\lim \bar{X}_n = \mu$.

The result in the previous Example is also known as the **weak law of large numbers** (WLLN). In order for the WLLN to apply the existence of the second moment (the variance) is not required. The WLLN holds just under the assumption that the $\{X_n\}_{n\in\mathbb{N}}$ $i.i.d.$ sequence have finite expected value μ.

Theorem

Consider two sequences $\{X_n\}_{n\in\mathbb{N}}$ and $\{Y_n\}_{n\in\mathbb{N}}$ of random variables converging in probability to $a<\infty$ and $b<\infty$, respectively. Then

1. $p\lim\left(X_n+Y_n\right)=p\lim X_n+p\lim Y_n=a+b.$
2. $p\lim\left(X_n\cdot Y_n\right)=p\lim X_n\cdot p\lim Y_n=a\cdot b.$
3. $b\neq 0,$
$$p\lim\left(\frac{X_n}{Y_n}\right)=\frac{p\lim X_n}{p\lim Y_n}=\frac{a}{b}.$$
4. Function g continuous in a : $p\lim g(X_n)=g\left(p\lim X_n\right)=g(a)$.

Definition

(Standardization) Let X be a random variable with expected value μ and finite variance σ^2. The location and scale trasform

$$Z = \frac{X - \mu}{\sigma}$$

defines the standardization of X. From the properties of expectation it is straightforward to prove that $Z \sim (0, 1)$.

Example

Let $\{X_n\}_{n \in \mathbb{N}}$ be an indipendent sequence of random variables with $X_n \sim (\mu, \sigma^2)$ for all $n \in \mathbb{N}$. Define the sequence $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$. Then

$$Z_n = \frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma} \sim (0, 1) \text{ for all } n \in \mathbb{N}.$$

Limit Theorems

Theorem

The Central Limit Theorem (CLT). Let $\{X_n\}_{n\in\mathbb{N}}$ *be independent random variables with a finite mean* $E[X_n] = \mu_n$ *and a finite variance* $Var(X_n) = \sigma_n^2$. *Define the normalized random variable*

$$Z_n = \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}}$$

so that $E[Z_n] = 0$ *and* $Var(Z_n) = 1$ *for all* n. *Then under regularity conditions the limiting distribution of* Z_n *is standard normal, denoted* $Z_n \to N(0,1)$, *i.e.*

$$\lim_{n\to\infty} F_{Z_n}(t) = \lim_{n\to\infty} P(Z_n \leq t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy.$$

Remark: the special condition X_n *independent with* $Var(X_n) = \sigma^2$ *for all* n *is sufficient for the CTL to apply.*

Limit Theorems

exercises ...

Parameter estimation

The object under study is

1. the probability distribution function F of a random experiment or random variable X, or
2. the statistical distribution function F of a given attribute of a population of individuals, users, devices, ...

We assume that F is known up to a vector of unknown parameters θ.

Definition

The family of distributions $\mathscr{P} = \{F_\theta\}_{\Theta \subseteq \mathbb{R}^n}$, $n \in \mathbb{N}$ finite, is called parametric model. The parametric model is usually specified in terms of probability mass or density functions.

Example

We assume that the number of e-mails per minute arriving to an e-mail server follows a Poisson distribution. The Poisson family of distributions is parametrized by a single parameter $\lambda > 0$

$$\mathscr{P} = \left\{ p_\lambda(j) = \frac{\lambda^j}{j!} \exp^{-\lambda}, \, j = 0, 1, \ldots \mid \lambda > 0 \right\}.$$

Example

The delivery time of the Google search engine is normal distributed. The Normal family is parametrized by two parameters $\theta = (\mu, \sigma)$

$$\mathscr{P} = \left\{ f_\theta(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp^{-\frac{1}{2\sigma^2}(x-\mu)^2} \mid \mu \in \mathbb{R}, \, \sigma > 0 \right\}.$$

Parameter estimation

Example

In the previous Example, instead of assuming the Normal distribution, we use the Logistic distribution, that is defined by the following distribution function (in this case $\theta = (\mu, \beta)$) :

$$\mathscr{P} = \left\{ F_\theta(x) = \frac{1}{1 + \exp^{-(x-\mu)/\beta}} \mid \mu \in \mathbb{R}, \beta > 0 \right\}.$$

Example

We are interested in the percentage of female students at USI. The distribution of the gender attribute of the USI population is Bernoulli distributed with parameter $\theta \in [0, 1]$.

In all previous examples the "true" vector θ of parameters identifying the correct distribution within the corresponding family is *unknown* and must be estimated by means of a sample (a random drawn subset of the population).

Parameter estimation

Parametric estimation theory deals with the problem of approximating the unknown parameters by means of the information collected from the sample.

Example

In order to estimate the percentage of female students we randomly select n students (sampling with replacement). In this way we obtain n independent realizations of Bernoulli distributed random variables X_i where

$$P(X_i = \text{"}female\text{"}) = \theta.$$

Identifying with 1 the female gender, the sample space is composed by 2^n points $x = (x_1, ..., x_n)$, where $x_i \in \{0, 1\}$ and $p_\theta(x) = \theta^{m(x)}(1-\theta)^{n-m(x)}$ with $m(x) = \sum_{i=1}^{n} x_i$. Given a sample x it seems reasonable to estimate θ by the proportion of successes i.e. $m(x)/n$.

Definition

The set of random variables $X_1, \ldots, X_n$ is said to constitute a *random sample* of size n from the population with the distribution function $F(x)$, provided that they are $i.i.d.$, i.e.

- mutually independent
- identically distributed

with distribution function $F_{X_i}(x) = F(x)$ for all i and all x.

Point estimation problem

The basic situation in point estimation is as follows.

- We observe a realization of random variables $X_1, \ldots, X_n$.
- The joint distribution function of $X_1, \ldots, X_n$ depends on an unknown parameter θ that is known to be in some given set Θ.
- The problem to find the value of θ is a problem of point estimation.
- We estimate θ by some function of the observations $x_1, \ldots, x_n$.

Definition

Any function of the random variables that are being observed, say $\hat{\Theta}(X_1, \ldots, X_n)$, is called a *statistic*. It is also called an *estimator of* θ. Since $X_1, \ldots, X_n$ are random variables, $\hat{\Theta}$ is a random variable too. A particular realization of the estimator, say $\hat{\Theta}(x_1, \ldots, x_n)$, is called an *estimate of* θ and denoted by $\hat{\theta}$.

Point estimation problem

Remark

- *The function $\hat{\Theta}$ must be known, that is, it does not have to depend on unknown parameters: $\hat{\Theta} = \overline{X}_n$ is a statistics, $\hat{\Theta} = \overline{X}_n - \mu$ is not a statistics if μ is unknown.*
- *The distribution of the estimator $\hat{\Theta}$ is called the sampling distribution.*
- *Since the joint distribution of $X_1, \ldots, X_n$ depends on θ, the sampling distribution will depends on θ too.*

Example

Let $X_1, \ldots, X_n$ be an $i.i.d.$ sequence of $N(\mu, 1)$ distributed random variables. From the properties of the Normal distribution we know that the sample distribution of $\hat{\Theta} = \overline{X}_n$ is $N(\mu, \frac{1}{n})$.

In many cases the sampling distribution is an unknown function of θ.

Point estimation problem

We would like our estimator of θ to be exactly θ on average.

Definition

We say that $\hat{\Theta}(X_1, ..., X_n)$ is an unbiased estimator of θ if

$$E\left[\hat{\Theta}(X_1, ..., X_n)\right] = \theta.$$

Example

The function

$$\hat{\Theta}(X_1, ..., X_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is an estimator of the population variance. It is possible to show (see Trivedi, page 642) that this estimator is biased, i.e.

$$E\left[\hat{\Theta}(X_1, ..., X_n)\right] = \sigma^2 - \frac{1}{n}\sigma^2 \neq \sigma^2.$$

Point estimation problem

Example

The formula

$$\hat{\Theta}(X_1, ..., X_n) = \sum_{i=1}^{n} a_i X_i$$

is an unbiased estimator of the expected value $\mu = E[X]$ provided that the real weights a_i satisfy the condition $\sum_{i=1}^{n} a_i = 1$.

- The previous example shows that many unbiased estimators of the same parameter may exist.
- We need a criterion to choose the "best" unbiased estimator.
- Recall that if $\hat{\Theta}$ is unbiased, then from chebyshev's inequality

$$P\left(|\hat{\Theta} - \theta| \geq \varepsilon\right) \leq \frac{Var[\hat{\Theta}]}{\varepsilon^2} \text{ for } \varepsilon > 0.$$

Point estimation problem

Definition

Efficiency. An estimator $\hat{\Theta}_1$ is said to be a more efficient estimator of the parameter θ than the estimator $\hat{\Theta}_2$, provided that

1. $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are both unbiased estimators of θ;
2. $Var[\hat{\Theta}_1] \leq Var[\hat{\Theta}_2]$, for all $\theta \in \Theta$;
3. $Var[\hat{\Theta}_1] < Var[\hat{\Theta}_2]$, for some $\theta \in \Theta$.

Example

The sample mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is the most efficient (minimum-variance) linear estimator of the population mean μ (provided μ exists). We also say that $\bar{X}$ is BLUE *(Best Linear Unbiased Estimator) for* μ (see Trivedi, page 644).

Point estimation problem

As in the case of the sample mean, the variance of the sampling distribution of an estimator generally decreases with increasing n. This fact leads us to the following property of an estimator.

Definition

Consistency. An estimator $\hat{\Theta}$ of the parameter θ is said to be consistent if $p\lim \hat{\Theta} = \theta$, i.e. if

$$\lim_{n \to \infty} P(|\hat{\Theta} - \theta| \geq \varepsilon) = 0.$$

From the Chebyshev inequality we conclude that any unbiased estimator $\hat{\Theta}$ of the parameter θ such that

$$\lim_{n \to \infty} Var[\hat{\Theta}] = 0$$

is consistent.

Point estimation problem

Example

The empirical distribution function,

$$\hat{F}(x) = \frac{\#\text{ observations } \leq x}{n}$$

is for all $x \in \mathbb{R}$ a consistent estimator of the true distribution function $F(x)$.
Proof: see Trivedi, page 645.

Method of least squares

Starting point:

- We observe a random sample y_1, ..., y_n from a population with distribution function F.
- We are interested in estimating the mean θ of the distribution.

We rewrite our model as

$$y_i = \theta + \varepsilon_i$$

where $\varepsilon_i := y_i - \theta$ is the zero mean deviation error of the $i - th$ observation from the mean of the underlying distribution. Note that $Var[\varepsilon_i] = Var[y_i] = \sigma^2$ for all observations. One possible criterion to estimate θ is to choose an estimate such that the sum of squared errors is as small as possible, i.e.

$$\hat{\theta}(y_1, ..., y_n) = \min_{\theta} \sum_{i=1}^{n} \varepsilon_i^2 = \min_{\theta} \sum_{i=1}^{n} (y_i - \theta)^2.$$

The solution is $\hat{\theta}(y_1, ..., y_n) = \bar{y}$ the mean of the sample!

We can extend the argument to the case where the expected value of the random variable Y_i depends on a concomitant **deterministic** and **observable** variable x_i. If the dependence is linear, we have that

$$E[Y_i] = \theta_1 + \theta_2 x_i,$$

where $\theta' = (\theta_1, \theta_2)$ is a vector of unknown parameters. As before, under this further assumptions we can write our model as

$$y_i = \theta_1 + \theta_2 x_i + \varepsilon_i$$

and estimate θ_1 and θ_2 by those number minimizing

$$\min_{\theta} \sum_{i=1}^{n} \varepsilon_i^2 = \min_{\theta} \sum_{i=1}^{n} (y_i - \theta_1 - \theta_2 x_i)^2.$$

Method of least squares

Finally, we can generalize the argument to the case of K explanatory variables $x_i' = (x_{i,1}, \ldots, x_{i,K})$. This means that, for the $i-th$ observation y_i, the model is

$$\begin{aligned} y_i &= \theta_1 x_{i,1} + \theta_2 x_{i,2} + \ldots + \theta_K x_{i,K} + \varepsilon_i \\ y_i &= x_i' \theta + \varepsilon_i \end{aligned}$$

We can rewrite the model in a more compact way using matrix notation

$$\underset{(n\times 1)}{y} = \underset{(n\times K)}{X} \underset{(K\times 1)}{\theta} + \underset{(n\times 1)}{\varepsilon}$$

where $y' = (y_1, \ldots, y_n)$, $\varepsilon' = (\varepsilon_1, \ldots, \varepsilon_n)$ and

$$X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,K} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,K} \end{bmatrix}.$$

Method of least squares

Again, the estimates of the K parameters $\theta_j, j = 1, ..., K$ are obtained by minimizing the sum of squared

$$\min_\theta \sum_{i=1}^n \varepsilon_i^2 = \min_\theta \varepsilon' \varepsilon = \min_\theta (y - X\theta)'(y - X\theta).$$

- This method of estimation is called the *method of least squares*.
- Any minimizing value $\hat{\theta}(y)$ is called the *least-square estimate* of θ.
- The function $\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^K$ is a *least-squares estimator*.

The solution of the minimization problem is obtained by differentiating with respect to θ the quadratic form $(y - X\theta)'(y - X\theta)$ and solving the first order conditions

$$X'X\theta = X'y.$$

Method of least squares

Any solution of the system of equations $X'X\theta = X'y$ is a least-square estimate. If rank $X = K$, then $X'X$ is non-singular and the unique least-square estimate is

$$\hat{\theta}(y) = (X'X)^{-1}X'y.$$

If rank $X < K$, then $X'X$ is singular and there is a family of least-square estimates.

Exercise

Assume the following regression model

$$y_i = \theta + \varepsilon_i \quad i = 1, \dots, n.$$

1. *Express the model in matrix form and give the dimension and the content of the matrix* X.
2. *Show that* $\hat{\theta}(y) = (X'X)^{-1}X'y = \bar{y}$.

Exercise

Assume the following regression model

$$y_i = \theta_1 + \theta_2 x_i + \varepsilon_i \quad i = 1, \ldots, n.$$

1. *Express the model in matrix form and give the dimension and the content of the matrix* X *in terms of the constant* 1 *and the explanatory variable* x_i, $i = 1, \ldots, n$.
2. *Show that* $\hat{\theta}_1 = \bar{y} - \hat{\theta}_2 \bar{x}$ *and* $\hat{\theta}_2 = \frac{\sum_i (x_i - \bar{x}) y_i}{\sum_i (x_i - \bar{x})^2}$.

Maximum likelihood estimation

Suppose that each night a certain lion has three possible states of activity:

1. very active, denoted by θ_1;
2. moderately active, denoted by θ_2;
3. lethargic, denoted by θ_3.

Each night this lion eats i people with probability $p(i|\theta)$, $\theta \in \Theta$ where $\Theta = \{\theta_j,\ j = 1, 2, 3\}$. The numerical values are given in the following table:

i	0	1	2	3
$p(i\|\theta_1)$	0.00	0.05	0.05	0.90
$p(i\|\theta_2)$	0.05	0.05	0.80	0.10
$p(i\|\theta_3)$	0.90	0.08	0.02	0.00

Question: If we know that $X = i_0$ people were eaten last night, how should we estimate the lion's activity state?

Maximum likelihood estimation

Answer: take as the estimate of θ that $\theta_j \in \Theta$ for which the probability $p(i_0|\theta)$ is largest! In this example we have $\hat{\theta}(i_0 = 0) = \theta_3$, $\hat{\theta}(i_0 = 1) = \theta_3$, $\hat{\theta}(i_0 = 2) = \theta_2$, $\hat{\theta}(i_0 = 3) = \theta_1$.

Suppose that we know for sure that $\theta \in \Theta' = \{\theta_1, \theta_2\}$. The estimate is no longer unique since it can be either θ_1 or θ_2.

Definition

Let $X_1, \ldots, X_n$ be n random variables with distribution function $F(x_1, \ldots, x_n|\theta)$ where $\theta \in \Theta$ is unknown. The *likelihood function* is

$$L(\theta) = \begin{cases} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n|\theta), & \text{if } F \text{ has a density function } f, \\ p_{X_1,\ldots,X_n}(x_1, \ldots, x_n|\theta), & \text{if } F \text{ has a prob. function } p. \end{cases}$$