Statistical measures & Distributions

Contents:

1. STATISTICAL MEASURES
   - (a) Mean
   - (b) Median
   - (c) Mode
   - (d) Range
   - (e) Standard Deviation
   - (f) Variance
2. DISTRIBUTIONS
   - (a) Binomial
   - (b) Poisson
   - (c) Normal
3. EXAMPLES

- The type of measures of MEAN, MEDIAN and MODE is called an *average* or *measure of location*.
- The type of measures of STANDARD DEVIATION and VARIANCE is called *measure of dispersion*.

The Mean

The ARITHMETIC MEAN (or just MEAN) of a set of numbers $\{x_1, x_2, \ldots, x_n\}$ is denoted by $\bar{x}$ and is defined by:

$$\boxed{\bar{x} = \frac{1}{n}(x_1 + x_2 \ldots x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.}$$

Consider a discrete frequency distribution taking values $\{x_1, x_2, \ldots, x_n\}$ with corresponding frequencies $\{f_1, f_2, \ldots, f_n\}$. The mean $\bar{x}$ is given by:

$$\boxed{\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}.}$$

Example

- Find the mean of the set $\{-3, -1, 0, 2, 3, 4\}$.
  $\bar{x} = (-3 - 1 + 0 + 2 + 3 + 4)/6 = 0.83$.
- Find the mean of the following frequency distribution:

x_i	-3	-2	-1	0	1	2	3
f_i	6	5	4	3	2	1	1
f_ix_i	-18	-10	-4	0	2	2	3

$\sum f_i = 22, \quad \sum f_i x_i = -25, \quad \bar{x} = -25/22 = -1.14.$

Let the set $\{x_1, x_2, \ldots, x_n\}$ be transformed to $\{X_1, X_2, \ldots, X_n\}$, where $X_i = (x_i - a)/b$. The $\bar{x}$ can be obtained by

$$\boxed{\bar{x} = a + b\bar{X}}$$

Proof

From $X_i = (x_i - a)/b$ it follows that $x_i = bX_i + a$. Now,

$$\begin{aligned} \bar{x} &= \frac{1}{n}\sum x_i = \frac{1}{n}\sum(a + bX_i) \\ &= \frac{1}{n}\sum a + \frac{b}{n}\sum X_i \\ &= a + b\bar{X}. \end{aligned}$$

□

Example

Find the mean of the set $\{2678, 4678, 8678, 5678, 6678\}$.

Let $X_i = (x_i - a)/b$, where $a = 2678$ and $b = 1000$. This gives the new set $X = \{0, 2, 6, 3, 4\}$ with mean $\bar{X} = 3$. Thus,

$$\bar{x} = 2678 + 1000 \times 3 = 5678$$

The Median

The mean has the disadvantage of taking extreme values into account, especially for a small set of numbers.

The MEDIAN of a set of numbers $\{x_1, x_2, \ldots, x_n\}$ is defined as the middled value of the set when arranged in size order. If the set has an even number of items, then the median is taken as the mean of the two middle two.

Example

1. The wages arranged in size order are: $\{28, 29, 32, 35, 36, 38, 41, 103\}$.

   The Mean is $\bar{x} = 41.89$ and the MEDIAN $x^* = 35.5$.

2. Find the median of the set: $\{65, 68, 68, 66, 64, 65, 65, 67\}$.

   Arranging the set in order: $\{64, 65, 65, 65, 66, 67, 68, 68\}$.

   The Median is given by: $(65 + 66)/2 = 65.5$.

Consider the discrete frequency distribution taking the values $\{x_1, x_2, \ldots, x_n\}$ with corresponding frequencies $\{f_1, f_2, \ldots, f_n\}$. The median is given by the

$$\boxed{\left(\frac{1+\sum f}{2}\right)\textbf{th}}$$

value when the values are ranked.

Example

Find the median of the following discrete distribution:

x_i	0	1	2	3	4	5	6
f_i	5	5	10	20	30	20	10
Cum f	5	10	20	40	70	90	100

$\sum f_i = 100, \quad (1+\sum f_i)/2 = 50.5.$

The 50.5th item falls at $x = 4$ using the Cumulative frequency (Cum f). Hence the *Median is* 4.

The Mode

The MODE of a set of values is defined as the one which occurs with the greatest frequency.

Note that for a set that has no repeated values the mode does not exist.

Example

The mode of the set $\{2,3,3,1,3,2,4,5,8,3,2,4,4,3\}$ is 3.

It should be noted that there should be more than one mode in a set of numbers.

Example

The set $\{8,6,8,5,5,7,6,8,6,9\}$ has the two modes 6 and 8.

The Range

The RANGE of a set of numbers $S = \{x_1, x_2, \ldots, x_n\}$ is given by:

$$\boxed{\textbf{Range} = \max(S) - \min(S).}$$

The Range is the simplest of all measures of dispersion and can be calculated very quickly and easily.

It is not a serious measure of dispersion since it uses the only extreme values.

Examples

- The set $\{6, \boxed{5}, 7, \boxed{10}, 8, 9\}$ has Range $= 10 - 5 = 5$.
- The set $\{600, 610, 620, \boxed{600}, 610, \boxed{650}, 640, 650, 650\}$ has Range $= 650 - 600 = 50$.
- The set $\{600, 610, 620, \boxed{200}, 610, \boxed{1000}, 640, 650, 650\}$ has Range $= 800$.

The Standard Deviation and Variance

The standard deviation is the measure of dispersion used most widely in statistics. It is based on the arithmetic mean.

The standard deviation of a set of numbers $\{x_1, x_2, \ldots, x_n\}$ with mean $\bar{x}$ is denoted by S and defined as

$$\begin{aligned} S &= \sqrt{\frac{\sum_i^n (x_i - \bar{x})^2}{n}} \\ &= \sqrt{\frac{1}{n}\sum_i^n x_i^2 - \bar{x}^2} \end{aligned}$$

Example

The set $\{3, 4, 6, 2\}$ has $\bar{x} = 15/4 = 3.75$, $\bar{x}^2 = 14.063$ and $\sum_i^n x_i^2 = 65$.

Thus,

$$\begin{aligned} S &= \left(\frac{65}{4} - (3.75)^2\right)^{\frac{1}{2}} = (16.25 - 14.063)^{\frac{1}{2}} \\ &= 1.48 \end{aligned}$$

For a discrete frequency distribution the standard deviation is defined as:

$$
\begin{aligned}
S &= \sqrt{\frac{\sum_i f_i(x_i-\bar{x})^2}{\sum_i f_i}} \\
&= \sqrt{\sum_i \frac{f_i x_i^2}{\sum_i f_i} - \bar{x}^2}
\end{aligned}
$$

where $\bar{x} = \sum_i f_i x_i / \sum_i f_i$ is the mean of the frequency distribution.

Example

Find the median of the following discrete distribution:

							Σ
x_i	0	1	2	3	4	5	
f_i	1	3	11	9	5	2	31
$f_i x_i$	0	3	22	27	20	10	82
x_i^2	0	1	4	9	16	25	
$f_i x_i^2$	0	3	44	81	80	50	258

$$
S = \sqrt{\sum_i \frac{f_i x_i^2}{\sum_i f_i} - \left(\frac{\sum_i f_i x_i}{\sum_i f_i}\right)^2} = \sqrt{\frac{258}{31} - \left(\frac{82}{31}\right)^2}
$$

$$
= 1.15
$$

The VARIANCE of a set (or distribution) of numbers is defined as the square of the standard deviation and is denoted by S^2.

For a set of numbers:

$$
\begin{aligned}
S^2 &= \frac{\sum_i^n (x_i-\bar{x})^2}{n} \\
&= \frac{1}{n}\sum_i^n x_i^2 - \bar{x}^2.
\end{aligned}
$$

For a frequency distribution:

$$
\begin{aligned}
S^2 &= \frac{\sum_i f_i(x_i-\bar{x})^2}{\sum_i f_i} \\
&= \sum_i \frac{f_i x_i^2}{\sum_i f_i} - \bar{x}^2.
\end{aligned}
$$

Let the set $\{x_1, x_2, \ldots, x_n\}$ be transformed to $\{X_1, X_2, \ldots, X_n\}$, where $X_i = (x_i - a)/b$. If the Standard deviations of X and x are denoted, respectively, by S_X and S_x, then

$$\boxed{S_x = bS_X.}$$

Proof

From $X_i = (x_i - a)/b$ it follows that $x_i = bX_i + a$. Now,

$$\begin{aligned} S_x &= \sqrt{\frac{(\sum x_i - \bar{x})^2}{n}} = \sqrt{\frac{(\sum bX_i + a - (a + b\bar{X}))^2}{n}} \\ &= b\sqrt{\frac{(\sum X_i - \bar{X})^2}{n}} \\ &= bS_X. \end{aligned}$$

□

Example

Find the standard deviation of the set $\{2678, 4678, 8678, 5678, 6678\}$.

Let $X_i = (x_i - a)/b$, where $a = 2678$ and $b = 1000$. This gives the new set $X = \{0, 2, 6, 3, 4\}$ with mean $\bar{X} = 3$ and standard deviation $S_X = \sqrt{20/5} = 2$. Thus,

$$S_x = bS_X = 1000 \times 2 = 2000.$$

Statistical measures & Distributions

Lecture 3

PROBABILITY DISTRIBUTIONS

1. Binomial
2. Poisson
3. Normal

Binomial Distribution

COMBINATIONS

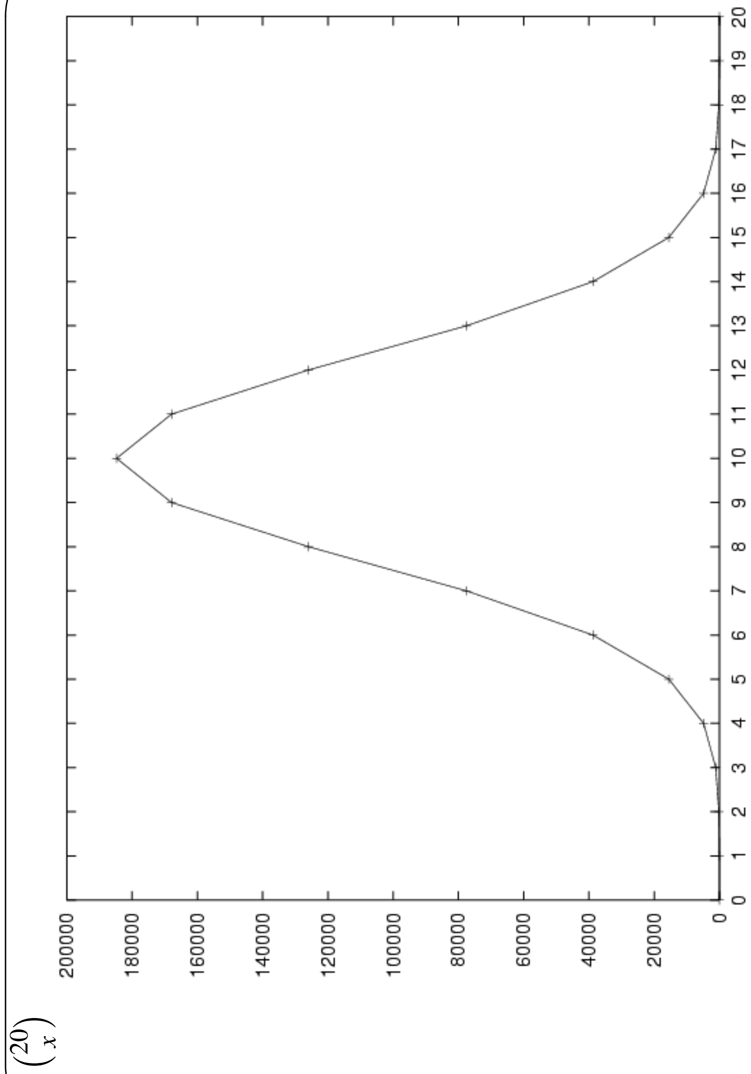
A COMBINATIONS is a non-ordered subset of a set of elements. Element order does not matter when determining Combinations.

The number of combinations of r elements taken from a set of n elements is given by:

$$\boxed{\binom{n}{r} = \frac{n\mathbf{!}}{r\mathbf{!}(n-r)\mathbf{!}} \quad \textbf{where} \quad r \leq n.}$$

The $n!$ (n factorial) is given by $n! = 1 \times 2 \times \cdots \times n$ and $0! = 1$.

The expression $\binom{n}{r}$ can be written as C_r^n, or ${}_nC_r$, or $C(n,r)$.



Example

Four assets designated A, B, C and D are considered by a fund manager. However he is only allowed to invest in three of them. How many different portfolios are possible?

Solution

The fund manager makes a selection of 3 assets from A, B, C and D without taking order in to account. He has 4 possible selections:

$$ABC, \quad ABD, \quad ACD, \quad BCD.$$

Otherwise, $n = 4$ and $r = 3$ so that

$$\binom{n}{r} = \binom{4}{3} = \frac{4!}{3!(1)!} = 4.$$

The BINOMIAL distribution arises in many applications where you are counting events.

The binomial distribution arises from the following assumptions:

1. It has a fixed number of trials, say n.
2. A random variable can take only two values (*success* or *failure*.)
3. Each trial is independent.
4. The probability of success, say p, is constant throughout the experiment.

Example

Consider an investor who is taking a buy position in 3 assets. If the price of the asset goes up (down), then the investor makes a profit (loss). Assume that the price distribution of the 3 assets included in the portfolio are independent. Now, let p ($q = 1 - p$) be the probability of the price going up (down). That is, $P(U) = p$ and $P(D) = q = 1 - p$. The probability of the first two assets going up and one going down is given by:

$$P(UUD) = P(U)P(U)P(D) = p\,p\,q = p^2\,q.$$

Note that $P(UUD) = P(UDU) = P(DUU)$.

Let $B(x;3;p)$ denote the probability that the prices of x assets rise. That is,

Port.	# of rises	Prob.
DDD	0	q^3
DDU	1	pq^2
DUD	1	pq^2
UDD	1	pq^2
DUU	2	p^2q
UDU	2	p^2q
UUD	2	p^2q
UUU	3	p^3

or

x	$B(x;3;p)$
0	q^3
1	$3pq^2$
2	$3p^2q$
3	p^3

or

$$B(x;3;p) = \binom{3}{x} p^x(1-p)^{3-x} \quad \textbf{where} \quad x = 0,1,2,3.$$

Note that
$\sum_1^3 B(x;3;p) = q^3 + 3pq^2 + 3p^2q + p^3 = (p+q)^3 = 1.$

The Binomial distribution has the form:

$$\boxed{P(X=x) = \binom{n}{x} p^x(1-p)^{n-x} \quad \textbf{for } x = 0,1,\ldots,n.}$$

Here the integer $n \geq 0$ and the probability p $(0 \leq p \leq 1)$ are the parameters.

The mean and variance of the Binomial distribution are given by:

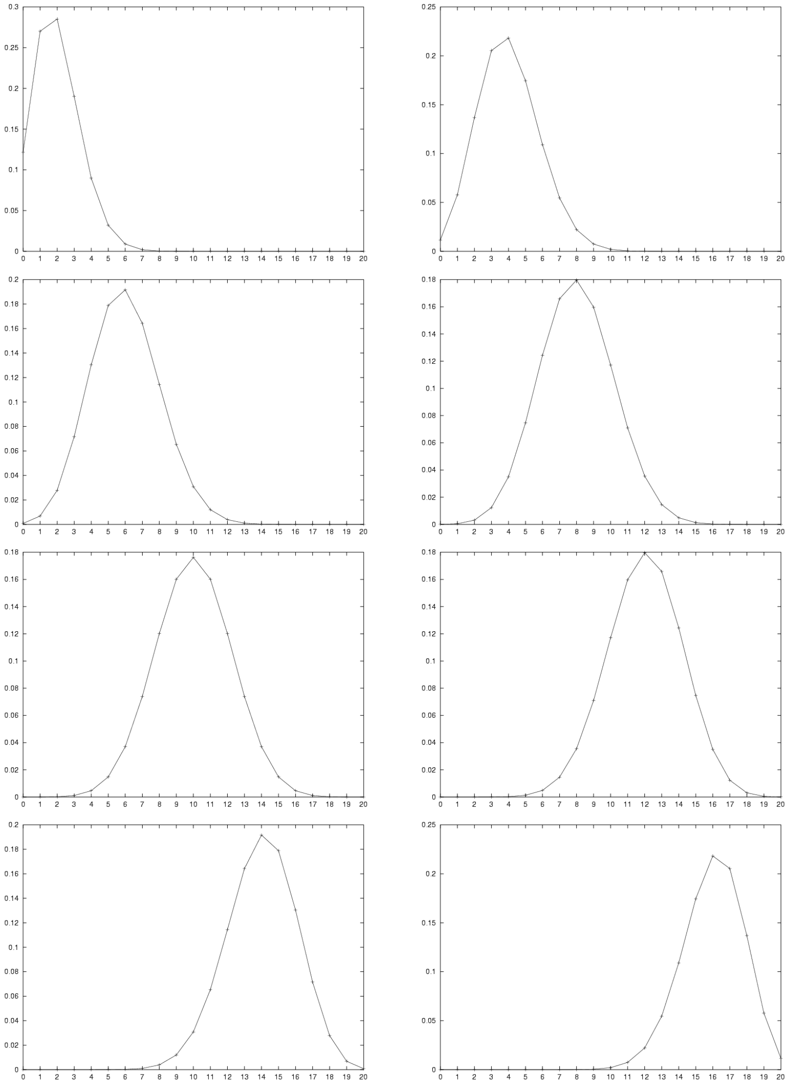
$$\mathbf{E}(x) = n\,p$$
$$\mathbf{Var}(x) = n\,p\,(1-p).$$

The relation

$$\boxed{P(X = x+1) = \frac{p}{1-p}\,\frac{n-x}{x+1}\,P(X=x)}$$

is known as the *Binomial recurrence formula*.

$B(x, 20, p)$, where $p = 0.1, 0.2, \ldots, 0.8, 0.9$.



Example

In tossing a coin the probability of a head is 0.5. If the coin is tossed 5 times, then what is the probability of (a) exactly 2 heads and (b) more than one head.

Let x denote the number of heads, $n = 5$ and $p = q = 1 - p = 1/2$.

1. $P(X = 2) = B(2; 5; p) = \binom{5}{2} p^2 (1-p)^3 = 5/16.$

2.

$$\begin{aligned} P(X > 1) &= 1 - P(X \leq 1) \\ &= 1 - \Big(P(x = 0) + P(x = 1)\Big) \\ &= 1 - B(0; 5; 0.5) - B(1; 5; 0.5) \\ &= 13/16. \end{aligned}$$

Given a frequency distribution the binomial distribution can be fitted by:

1. Deriving the values of the parameters n and p either
   (a) From a know binomial situation
   (b) Computing the mean $\bar{x}$ of the frequency distribution and using the relation $\bar{x} = np$.
2. Generating a Binomial probability distribution using n and p.
3. Generate the *Expected* frequencies by multiplying the total frequencies $\sum_i f_i$ by the probability.

Example

A *biased* die is thrown 5 times as an experiment. The experiment is repeated 250 times. The number of even numbers shown on the die in each experiment is recorded giving the results:

# of *evens*	0	1	2	3	4	5	Total
Observed freq.	11	41	83	73	36	6	250

FIT A BINOMIAL DISTRIBUTION TO THIS DATA.

Step 1

The number of trials $n = 5$. The relationship $\bar{x} = np$ can be used in order to compute p.

Now,

x_i	0	1	2	3	4	5	Total
f_i	11	41	83	73	36	6	250
$f_i x_i$	0	41	166	219	144	30	600

It follows that $\bar{x} = \sum_i f_i x_i / \sum_i f_i = 600/250 = 2.4$.

Thus, from $\bar{x} = np$ it implies $2.4 = 5\,p$, i.e. $p = 0.48$.

Step 2

Let the random variable X which represent the *number of evens in the experiment* to have the binomial distribution:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \textbf{for } x = 0, 1, 2, 3, 4, 5,$$

where $n = 5$ and $p = 0.48$.

Now $P(X=0)=\binom{5}{0}p^0(1-p)^5=0.038$ and

$$\begin{aligned} P(X=x+1) &= \frac{p}{1-p}\,\frac{n-x}{x+1}\,P(X=x) \\ &= 0.9231\,\frac{5-x}{x+1}\,P(X=x). \end{aligned}$$

From the latter it follows that $P(X=1)=0.1754$, $P(X=2)=0.324$, $P(X=3)=0.2990$, $P(X=4)=0.1380$ and $P(X=5)=0.0255$.

Summary

x_i**:** The number of *evens*.

f_i**:** Actual frequency.

$P(X=x_i)$**:** Probability of obtaining x_i *evens*.

$P(X=x_i)\sum_i f_i$**:** Expected frequency.

x_i	0	1	2	3	4	5	Total
f_i	11	41	83	73	36	6	250
$P(X=x_i)$	0.04	0.18	0.32	0.3	0.14	0.03	0.9999
$P(X=x_i)\sum_i f_i$	10	44	81	75	34	6	250

Poisson Distribution

The discrete random variable X is said to have a Poisson distribution if it has a pdf of the form

$$\boxed{P(X=x)=\frac{e^{-\lambda}\lambda^x}{x!} \qquad \textbf{for } x=0,1,\ldots,n,}$$

where $\lambda>0$ is the parameter.

The mean and variance of the Poisson distribution are given by:

$$\boxed{\mathbf{E}(x)=\mathbf{Var}(x)=\lambda.}$$

The Poisson distribution can be defined as the limiting case of the binomial distribution for $n\to\infty$ but with constant $np=\lambda$. Thus, it describes the behavior of a large number n of independent experiments of which only a very small fraction np is expected to yield events of a given type.

Example

The claim experience of 5000 policies, each expose to risk for a year, is summarized in the table below:

x_i	0	1	2	3	Total
f_i	3695	1120	160	25	5000

Here x_i is the number of claims and f_i the observed number of policies.

Questions

1. Calculate the average number of claims.
2. Assume a Poisson distribution. Compare the predictions of the theoretical distribution with the observed number of policies.

Answer

1. The average number of claims is calculated by:

$$\lambda = \frac{\sum_i f_i x_i}{\sum_i f_i} = \frac{1515}{5000} = 0.303$$

2. First calculate $P(X = x)$ for $x = 0, 1, 2, 3$.

- $P(X=0) = \dfrac{e^{-\lambda}\lambda^0}{0!} = e^{-\lambda} = 0.7386.$
- $P(X=1) = \dfrac{e^{-\lambda}\lambda^1}{1!} = P(X=0) \times \dfrac{\lambda}{1} = 0.2238.$
- $P(X=2) = \dfrac{e^{-\lambda}\lambda^2}{2!} = P(X=1) \times \dfrac{\lambda^2}{2} = 0.0339.$
- $P(X=3) = P(X=2) \times \dfrac{\lambda^3}{3} = 0.0034.$

Now the predicted number of policies for x claims is given by $\hat{f}_i = P(X = x_i)\sum_i f_i$. Thus, $\hat{f}_1 = P(X = x_i)\sum_i f_i = 0.7386 \times 5000 = 3693.$

Number	Probab. of claims	Number of policies	
of claims	per policy	Predicted	Actual
x_i	p_i	$\hat{f}_i$	f_i
0	0.7386	3693.0	3695
1	0.2238	1119.0	1120
2	0.0339	169.5	160
3	0.0034	17.0	25
TOTAL	0.9997	4998.5	5000

Comparing the theoretical with the empirical values we observe that a Poisson distribution with parameters value of $\lambda = 0.3$ describes this particular random variable very well.

Observations

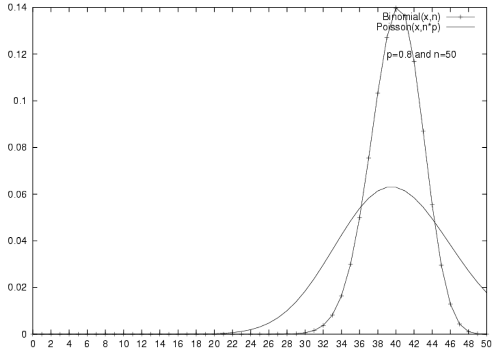
1. $\sum_i p_i = 0.9997 \approx 1$.
2. $\mathrm{E}(x) = \sum_i x_i p_i = 0.3018 \equiv \lambda$.
3. $\mathrm{E}(x^2) = \sum_i x_i^2 p_i = 0.39$.
4. $\mathrm{Var}(x) = \mathrm{E}(x^2) - (\mathrm{E}(x))^2 = 0.299 \approx \lambda$. .

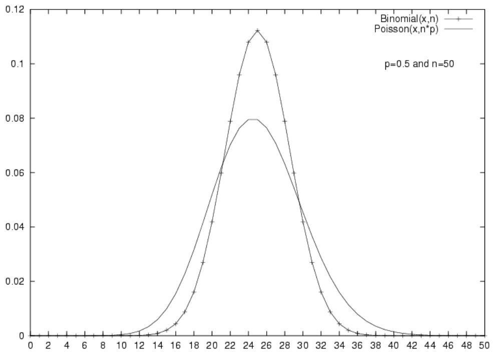
Binomial approximation to Poisson distribution

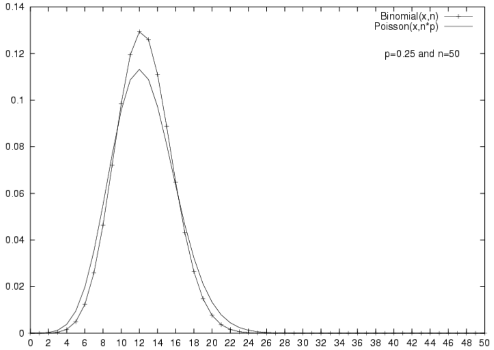
A Binomial distribution with parameters n and p can be approximated by a Poisson distribution with parameter $\lambda = n\,p$ if n is large and p is small. That is, $n \to \infty$ and $p \to 0$.

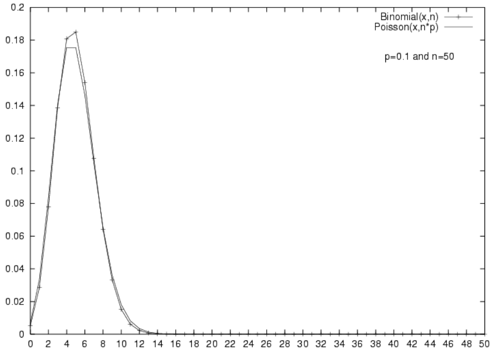
Note that if $p \to 0$, then $q = 1 - p \approx 1$. Thus, the variance of the Binomial distribution is given by:

$$\boxed{\mathbf{Var}(x) = n\,p\,(1-p) \approx n\,p\,1 = n\,p = \mathbf{E}(x).}$$









Example

Consider an individual who has insured his car against theft. The probability of a theft in any 24-hour period, leading to an insurance claim is 0.005. The probabilities of claims on successive days are independent. In addition it is not possible to have more than one theft (leading to a claim) on the same day.

Calculate the probability that a policyholder makes at least 3 claims in a year.

The probability of theft is $p = 0.005$. The probability of the car not to be stolen (i.e. No theft) is $q = 1 - p = 0.995$. The total number of trials in a year is the number of days, that is, $n = 365$.

The probability of at least 3 claims is given by:

$$\begin{aligned} P(X \geq 3) &= 1 - P(X \leq 2) \\ &= 1 - \Big(P(X = 0) + P(X = 1) + P(X = 2)\Big). \end{aligned}$$

Using the Binomial distribution and recurrence formula:

$$P(X=0) = \binom{365}{0} p^0 q^{365} = 0.1605,$$

$$P(X=1) = 0.2944 \quad \textbf{and}$$

$$P(X=2) = 0.2692.$$

$$\textbf{Thus,} \quad P(X \geq 3) = 0.276\textbf{.}$$

Now, Using an approximation to the Poisson Distribution. Let $\lambda = n\,p = 365 \times 0.005 = 1.825$. Note that $npq = 1.815$.

$$P(X=0) = \frac{e^{-\lambda}\lambda^0}{0\textbf{!}} = 0.1612,$$

$$P(X=1) = 0.2942 \quad \textbf{and}$$

$$P(X=2) = 0.2685.$$

$$\textbf{Thus,} \quad P(X \geq 3) = 0.2761\textbf{.}$$

The difference using a Poisson distribution has relative error 0.04% which is extremely small.

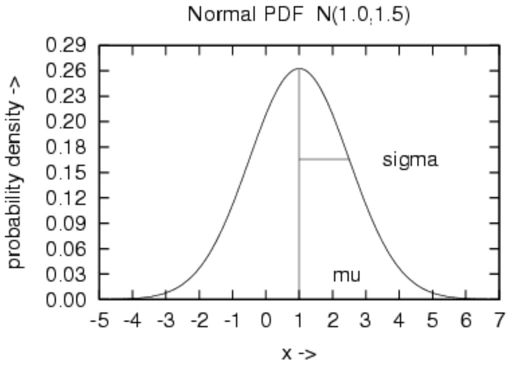
The Normal Distribution

The Binomial and Poisson distributions were important examples of special distributions of the discrete kind. The NORMAL distribution can be described as the single most important continues distribution in statistics.

The Normal distribution has two parameters: the mean μ and the standard deviation σ. Its shorthand notation is $N(\mu,\sigma)$. The pdf of $N(\mu,\sigma)$ is bell shaped and is symmetrical about the mean. The formula is:

$$\boxed{f(x) = \frac{1}{\sigma\sqrt{2\Pi}} e^{-\frac{1}{2}\left(\frac{(x-\mu)}{\sigma}\right)^2} \quad \textbf{for } -\infty < x < \infty\textbf{,}}$$

where $\mathrm{E}(x) = \mu$ and $\mathrm{Var}(x) = \sigma^2$.



A normal distribution having $\mu = 0$ and $\sigma^2 = 1$, i.e. $N(0,1)$, is called a STANDARD NORMAL DISTRIBUTION. The random variable associated with this distribution is usually denoted by Z. That is, $Z \sim N(0,1)$. The pdf of Z is given by:

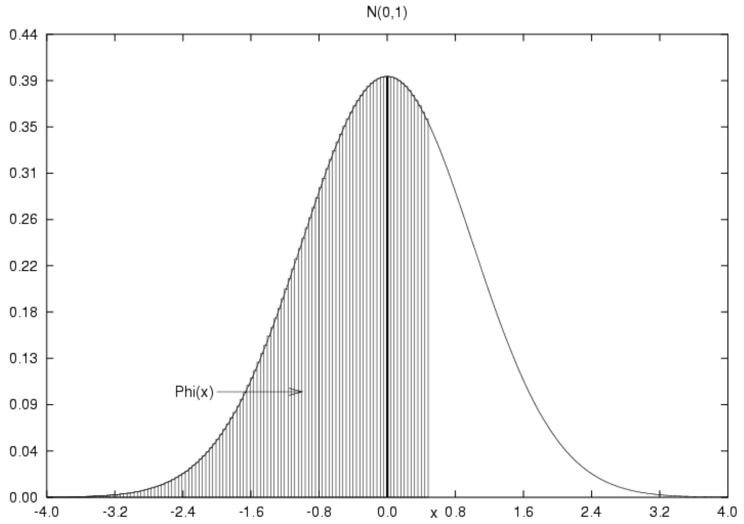
$$\boxed{f(x) = \frac{1}{\sqrt{2\Pi}} e^{-\dfrac{x^2}{2}} \quad \textbf{for } -\infty < x < \infty\textbf{.}}$$

The distribution function of a standard normal variable Z is denoted by:

$$\Phi(x) = P(Z < x) = \int_{-\infty}^{x} f(x)\, dx$$

$$= \frac{1}{\sqrt{2\Pi}} \int_{-\infty}^{x} e^{-\dfrac{x^2}{2}}\, dx.$$

Close form expression for the integral does not exist. Hence its evaluation can only be obtained by approximate procedures. Therefore, areas under the normal density function are presented in tables.



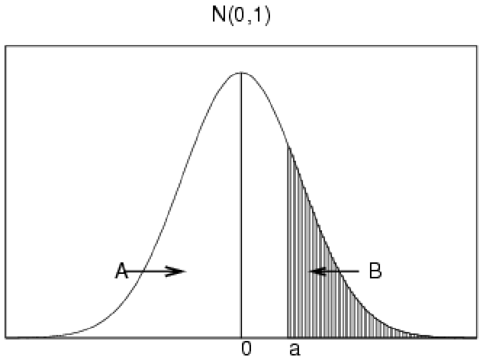
Example

$P(Z < 0.1) = \Phi(0.1) = 0.5398.$

The Normal distribution tables only give values of $\Phi(x)$ for $x \geq 0$. The probabilities such as $P(Z < -0.1)$ and $P(Z \geq 0.3)$ have to be *transformed* into probabilities of the type $P(Z < x)$, where $x \geq 0$.

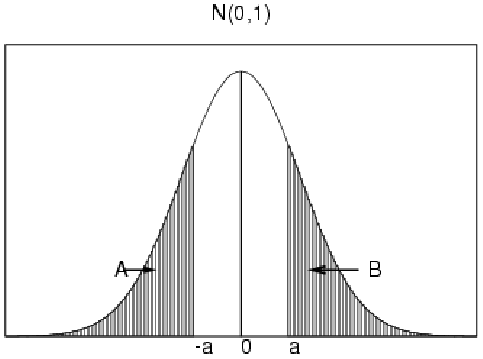
Let $Z \sim N(0,1)$ and $\Phi(x) = P(Z < x)$. If $a \geq 0$, then

1. $P(Z > a) = 1 - P(Z < a) = 1 - \Phi(a)$.



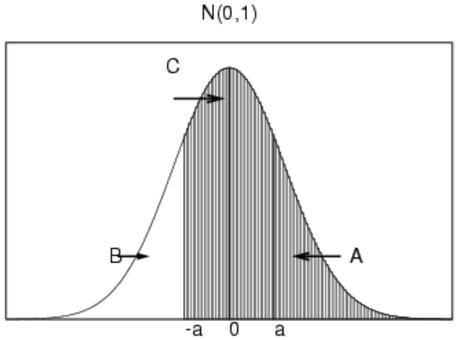
$P(Z > a) =$ **Area(B)** $=$ **Total Area - Area(A)** $= 1 - \Phi(a)$.

2. $P(Z < -a) = \Phi(-a) = 1 - \Phi(a)$.



$\Phi(-a) =$ **Area(A)** $=$ **Area(B)** $= P(Z > a) = 1 - \Phi(a)$.

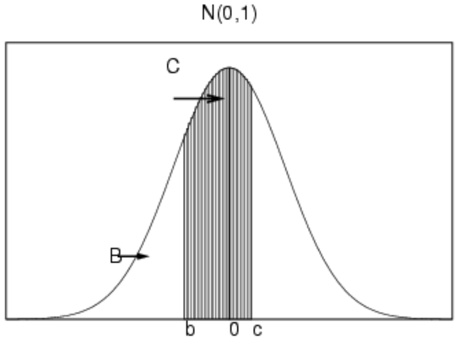
3. $P(Z > -a) = \Phi(a)$.



$P(Z > -a) =$ **Area(C)** $+$ **Area(A)**

$=$ **Area(C)** $+$ **Area(B)** $= P(Z < a) = \Phi(a)$.

4. If b and c are any positive or negative numbers such that $b \leq c$, then $P(b < Z < c) = \Phi(c) - \Phi(b)$.



$P(b < Z < c) =$ **Area(C)** $=$ **Area(B+C)** $-$ **Area(B)**

$= P(Z < c) - P(Z < b) = \Phi(c) - \Phi(b)$.

Example 1

The random variable $X \sim N(\mu, \sigma^2)$ denotes the number of claims per year, where $\mu = 100$ and $\sigma = 4$. Find the probabilities of the number of claims to be:

1. Less that 90.
2. More than 108.
3. Between 96 and 104 (including).

Solution

1.

$$P(X < 90) = P\Big(\frac{X-100}{4} < \frac{90-100}{4}\Big)$$
$$P(Z < -2.5) \quad \textbf{since} \quad (X-\mu)/\sigma \sim N(0,1)$$
$$1 - \Phi(2.5) = 0.0062$$

2.

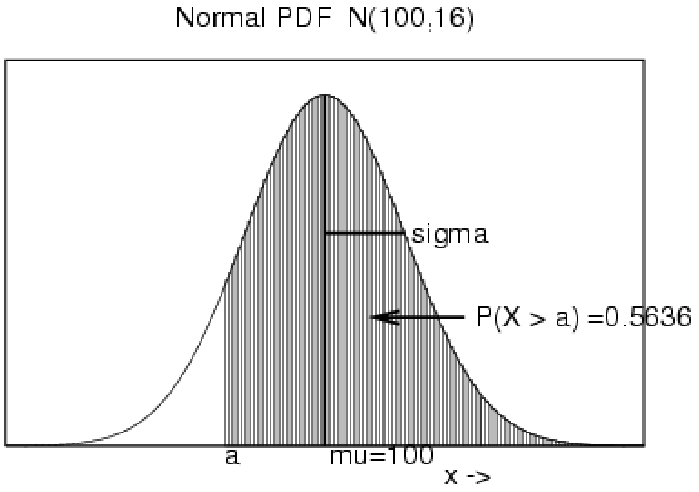
$$P(X > 108) = P\Big(\frac{X-100}{4} > \frac{108-100}{4}\Big)$$
$$P(Z > 2.0) = 1 - \Phi(2.0)$$
$$1 - 0.9772 = 0.0228.$$

3.

$$\begin{aligned} P(95 < X < 105) &= P\Big(\frac{95-100}{4} < \frac{X-100}{4} < \frac{105-100}{4}\Big) \\ &= P(-1.25 < Z < 1.25) \\ &= \Phi(1.25) - \Phi(-1.25) \\ &= 2 \times \Phi(1.25) - 1 = 0.7888. \end{aligned}$$

Example 2

The probability of having more than a claims is 0.5636. What is the value of a, when $X \sim N(100, 16)$.



Since the probability given is greater than 0.5 then a must be less than the mean $\mu = 100$.

Now, $P(X > a) = 0.5636$ and thus,

$$P\Big(\frac{X-100}{4} > \frac{a-100}{4}\Big) = 0.5636$$
$$\textbf{or} \quad P\Big(Z > \frac{a-100}{4}\Big) = 0.5636.$$

Since a is less than the mean, then $(a-100)/4 < 0$. Therrefore,

$$P\Big(Z > \frac{a-100}{4}\Big) = P\Big(Z > -(\frac{100-a}{4})\Big)$$
$$= \Phi\Big(\frac{100-a}{4}\Big).$$

Hence,

$$\Phi\Big(\frac{100-a}{4}\Big) = 0.5636$$

From the tables it follows that

$$\frac{100-a}{4} = 0.16 \quad \textbf{and thus,} \quad a = 99.36\textbf{.}$$

Example 3

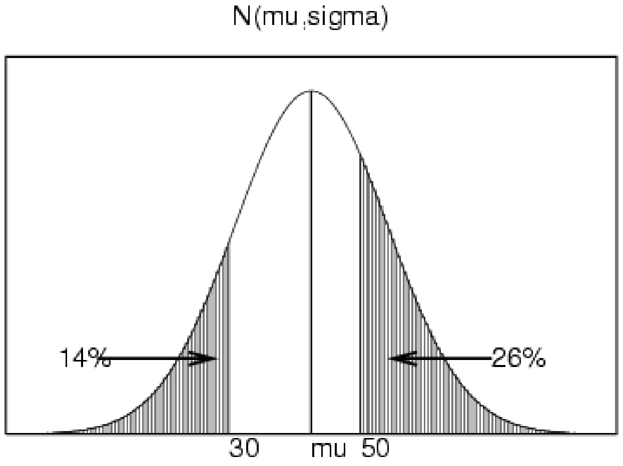
A policyholder has a large number of policies. If the policies are known to be normally distributed and

1. 14% of the policies gave rise to less than 30 claims;
2. 26% gave rise to more than 50 claims.

FIND THE MEAN AND VARIANCE OF THE CLAIMS.

Answer

Let X denotes the claims, i.e. $X \sim N(\mu, \sigma^2)$. The information given by (1) and (2) can be written as $P(X < 30) = 0.14$ and $P(X > 50) = 0.26$. Graphically this can be illustrated as:



Now, $P(X < 30) = 0.14$ can equivalently be written as

$$P\Big(\frac{X-\mu}{\sigma} < \frac{30-\mu}{\sigma}\Big) = 0.14$$

$$\textbf{or} \quad P\Big(Z < -(\frac{\mu-30}{\sigma})\Big) = 0.14$$

$$\textbf{or} \quad \Phi\Big(-(\frac{\mu-30}{\sigma})\Big) = 0.14$$

$$\textbf{or} \quad 1-\Phi\Big(\frac{\mu-30}{\sigma}\Big) = 0.14 \quad \textbf{or} \quad \Phi\Big(\frac{\mu-30}{\sigma}\Big) = 0.86.$$

From the tables it follows that $\Phi(1.08) = 0.8599$. Thus,

$$\frac{\mu-30}{\sigma} = 1.08 \quad \textbf{or} \quad \mu - 1.08\times\sigma = 30. \qquad (1)$$

Clearly, from $P(X > 50) = 0.26$ it follows that

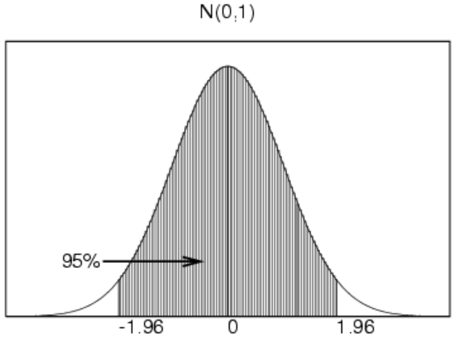
$$P\Big(Z > \frac{50-\mu}{\sigma}\Big) = 0.26, \quad \textbf{where} \quad Z = \frac{X-\mu}{\sigma}.$$

Thus,
$1-\Phi\Big(\frac{50-\mu}{\sigma}\Big) = 0.26$ or $\Phi\Big(\frac{50-\mu}{\sigma}\Big) = 0.74$. From the tables it follows that

$$\frac{50-\mu}{\sigma} = 0.643 \quad \textbf{or} \quad \mu + 0.643\times\sigma = 50. \qquad (2)$$

From (1) and (2) it follows that the mean $\mu = 42.54$ and variance $\sigma^2 = (11.61)^2 = 134.8$.

The central 95% of a standard normal distribution lies between the limits ±1.96. Alternatively, the central 95% of any normal distribution lies within 1.96 standard deviations of its means.



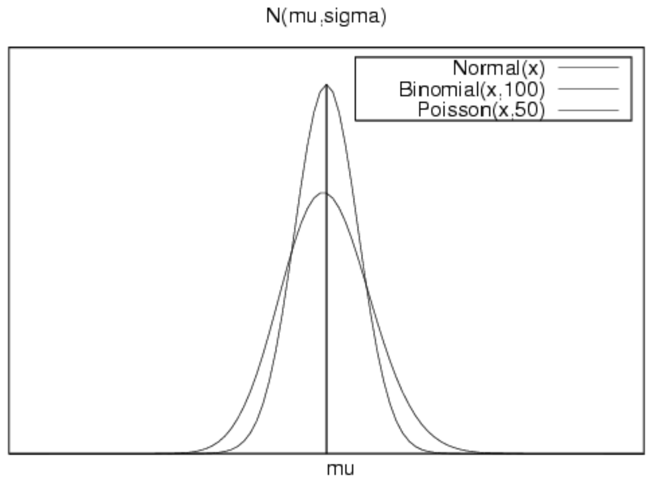
The central 99% (99.8%) of a standard normal distribution lies between the limits ±2.58 (±3.09).

Normal Approximation to the Binomial and Poisson

- If X is distributed binomially with parameters n and p, i.e. $X \sim Bin(n,p)$, then for large n and not too small (or too large) p we can consider $X \sim N\big(np, np(1-p)\big)$.

  The number of trial n should, in general, $n > 50$ and $p \approx 0.5$.
- If X has a Poisson distribution with parameter μ, i.e. $X \sim Po(\mu)$, then for large μ, we canconsider $X \sim N(\mu,\mu)$.

  In general the parameter μ>20 for good approximations.



Central limit theorem

If $X_1, X_2, \ldots, X_n$ is a random sample of size n from *any distribution* with mean μ and variance σ^2, then the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ has an approximate normal distribution with mean μ and variance σ^2/n. That is, approximately

$$\boxed{\bar{X} \sim N(\mu, \sigma^2/n).}$$

The approximation gets better for large n.

Example

Let $X_1, X_2, \ldots, X_n$ denote a random sample of claims from Poisson distribution with parameter $\mu = 3$, where $n = 20$. Using the central limit theorem find the approximation that the sample mean will be greater than 4.

Solution

If X is Poisson with parameter 3, then $\mathrm{E}(X) = \mu = 3$ and $\mathrm{Var}(X) = \sigma^2 = \mu = 3$. From the Central Limit Theorem it follows that $\bar{X} \sim N(\mu, \mu/n) = N(3, 0.15)$ approximately.

Hence,

$$
\begin{aligned}
P(\bar{X} > 4) &= P\Big(\frac{\bar{X}-3}{\sqrt{0.15}} > \frac{4-3}{\sqrt{0.15}}\Big) \\
&= P(Z > 2.58) \quad (\text{where } Z = (\bar{X}-3)/\sqrt{0.15}) \\
&= 1 - \Phi(2.58) = 0.005.
\end{aligned}
$$

Example

A population of insured clients has a mean claim (in pounds) of $\mu = 69$ and a standard deviation $\sigma = 3.22$. If a random sample of $n = 10$ insured is drawn, then what is the chance that the mean $\bar{X}$ will be within £2 of the population mean μ ?

Solution

We want to find

$$
\begin{aligned}
P(|\bar{X}-\mu| < 2) &= P(-2 < \bar{X}-\mu < 2) \\
&= P(67 < \bar{X} < 71) \\
&= 1 - P(\bar{X} < 67) - P(\bar{X} > 71).
\end{aligned}
$$

Now, according to the central limit theorem, $\bar{X} \sim N(\mu, \sigma^2/n) = N(69, 1.04)$.

Let $Z = (\bar{X}-69)/\sqrt{1.02}$, such that after standardization the $P(|\bar{X}-\mu| < 2)$ can be written as

$$
\begin{aligned}
P(|\bar{X}-\mu| < 2) &= 1 - P(Z < -1.96) - P(Z > 1.96) \\
&= 1 - 0.025 - 0.025 \\
&= 0.95.
\end{aligned}
$$

Thus, there is a 95% chance that the sample mean will be within £2 of the population mean.

Example

Assume that a large class in quantitative methods has marks normally distributed around mean of 72 with a standard deviation of 9.

1. Find the probability that an individual student drawn at random will have a mark over 80.
2. Find the probability that a random sample of 10 students will have an average mark over 80. What will be this probability when the population is NOT normal.

Solution

1. Let X denote the marks of the students, where $\mu = 72$ and $\sigma^2 = 81$. Thus, $X \sim N(72, 81)$.

   We want the probability $P(X > 80)$.

   Let $Z = (X - 72)/9$. Notice that $(80 - 72)/9 = 0.89$. Thus, after standardizing $P(X > 80) = P(Z > 0.89) = 0.187$.

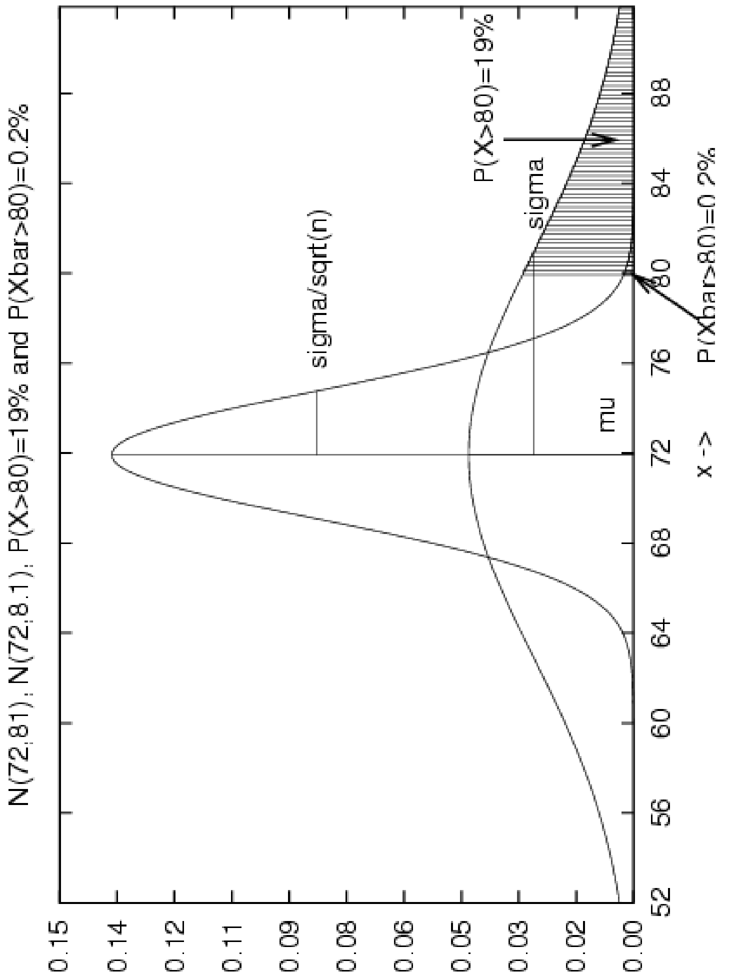
2. From the central limit theorem we have that $\bar{X} \sim N(\mu, \sigma^2/n) = N(72, 8.1)$, where $\bar{X}$ denotes the average mark of a random sample of 10 students.

   We want the $P(\bar{X} > 80)$.

   Now let, $Z = (\bar{X} - 72)/2.85$ and notice that $(80 - 72)/2.85 = 2.81$. Thus, after standardizing:

   $$P(\bar{X} > 80) = P(Z > 2.81) = 0.002.$$

   The CLT implies that for sufficient large n the $\bar{X} \sim N(\mu, \sigma^2/n)$ (approximately) no matter what the distribution of the parent population. Thus, $P(\bar{X} > 80)$ is approximately 0.2%.



Interval estimation

The purpose of interval estimation is to construct ranges of values within which the population parameters are expected to lie within a given probability based on the results of a random sample.

A *B% confidence interval* (C.I.) for some unknown parameter θ is an interval constructed based on the results of a random sample so that the probability θ lies in this interval is $B/100$.

The most commonly used is a 95% C.I. If (a,b) constitute a 95% C.I. for some parameter θ, then we have in probability terms: $P(a \leq \theta \leq b) = 0.95$.

The construction of the intervals that we consider will be based on the sample values of unbiased estimators for the particular parameters that we are interested.

The mean (known variance)

If $\bar{X}$ is the mean of a random sample of size n from a normal distribution with known variance σ^2, the a central 95% C.I. for μ, the population mean, is given by $\bar{X} \pm 1.96\sigma/\sqrt{n}$. That is,

$$\boxed{P\Big(\bar{X} - \frac{1.96\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{1.96\sigma}{\sqrt{n}}\Big) = 0.95.} \qquad (3)$$

Exercise

An experiment was carried out in which it was found that the height in cm of a plan were

$$\{12.3, 11.8, 11.6, 12.6, 13.4, 12.8, 11.1, 12.2, 14.8, 13.1\}$$

Given that the height of the plans are approximately normally distributed with variance $1.44cm$, construct a 95% C.I. for the mean of the population heights μ.

Solution

$P(11.83 \leq \mu \leq 13.31) = 0.95.$

The mean with unknown variance (small samples)

Let $\bar{X}$ and S^2 denote the mean and variance of a random sample of size n drawn from a normal population with unknown mean μ and unknown variance σ^2. The central $B\%$ C.I. for μ is given by $\bar{X} \pm t\, S/\sqrt{n-1}$, where t is such that the interval $(-t, t)$ encloses $B\%$ of a $T(n-1)$ distribution. That is,

$$\boxed{P\Big(\bar{X} - t\,\frac{S}{\sqrt{n-1}} \le \mu \le \bar{X} + t\,\frac{S}{\sqrt{n-1}}\Big).} \qquad (4)$$

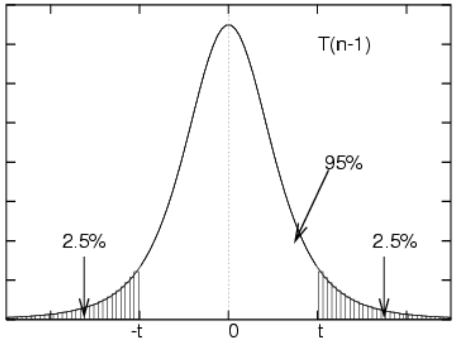
Note that the latter form is also used for samples from populations that are approximately normal.

Example

Eight shares of a certain PLC are reported to be (in £) $\{10.6, 11.2, 10.4, 12.2, 11.3, 10.2, 10.3, 12.5\}$.

Find a 95% confidence limits for the mean price of a sample of the PLC assuming these prices are coming from a normal distribution.

Solution



The confidence limits take the form $\bar{X} \pm t\, S/\sqrt{n-1}$, where $(-t, t)$ is the interval of a $T(n-1)$ distribution enclosing the central 95% of the distribution. From T tables with $n = 7$ we have that $t = 2.365$. That is, $T_{2.5\%}(7) = 2.365$.

From the share prices we have $\bar{X} = (\sum X_i)/n = 88.7/8 = 11.09$ and $\sum X_i^2 = 988.7$. Thus, $S^2 = \frac{1}{n}\sum X_i^2 - \bar{X}^2 = 0.68$ and $S = 0.82$.

From the latter it follows that 95% C.I. is

$$11.09 \pm \frac{2.365 \times 0.82}{\sqrt{7}} = 11.09 \pm 0.73 = (10.36, 11.82).$$

Thus, we have a probability of 95% that the mean share price of the PLC will lie between 10.36 and 11.82.

The mean with unknown variance (Large sample approx.)

For large values of n, the $T(n)$ closely resembles the standard Normal distribution, i.e. $N(0,1)$. For example $Z_{2.5\%} = 1.96$ which implies that $(-1.96, 1.96)$ encloses the central 95% of a $N(0,1)$ distribution. Similarly, from column $P = 2.5\%$ and $n = 120$ of T-tables we obtain $t = 1.98$. That is, the central 95% of a $T(120)$ distribution lies within the interval $(-1.98, 1.98)$.

Let $\bar{X}$ and S^2 denote the mean and variance of a random sample of size n (large) from a normal population with unknown mean μ and unknown variance σ^2. A central $B\%$ confidence interval for μ is given (approximately) by $\bar{X} \pm zS/\sqrt{n}$, where z is the $0.5(100-B)\%$ point of a $N(0,1)$ distribution. That is,

$$\boxed{P\Big(\bar{X} - z\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + z\frac{S}{\sqrt{n}}\Big) = \frac{B}{100}.}$$

Note that $n \geq 30$ will be considered adequately large.

Example

One hundred shares from the same kind of business are taken at random and their prices are found to have mean 69 and variance 7. Find 95% and 98% confidence limits for the mean price of the shares.

Solution

From the large sample $n = 100$ we have the mean $\bar{X} = 69$ and variance $S^2 = 7$. Since n is large we compute $\bar{X} \pm zS/\sqrt{n}$. For 95% C.I., the $z = 1.96$ and gives the limits as: $69 \pm 1.96(\sqrt{7/100}) = 69 \pm 0.52$.

That is, the 95% C.I. is given by $(68.48, 69.52)$.

For 98% C.I., the $z = 2.33$. Thus, ...

Regression

Contents:

1. SIMPLE LINEAR REGRESSION
2. MULTIPLE REGRESSION
3. REGRESSION DIAGNOSTICS

Simple regression

In practice we often want to study more than one variable. We usually want to look at how one variable is related to other variables.

Regression analysis is used for explaining or modelling the relationship between a single variable y, called the *response*, *output* or *dependent* variable, and one or more *predictor*, *input*, *independent*, or *explanatory* variables $x_1, x_2, \ldots, x_n$. If $n = 1$, then it is called simple regression; otherwise, if $n > 1$ it is called multiple regression, or sometimes multivariate regression. When there are more than one y, then it is called multivariate multiple regression.

Regression has several possible objectives including:

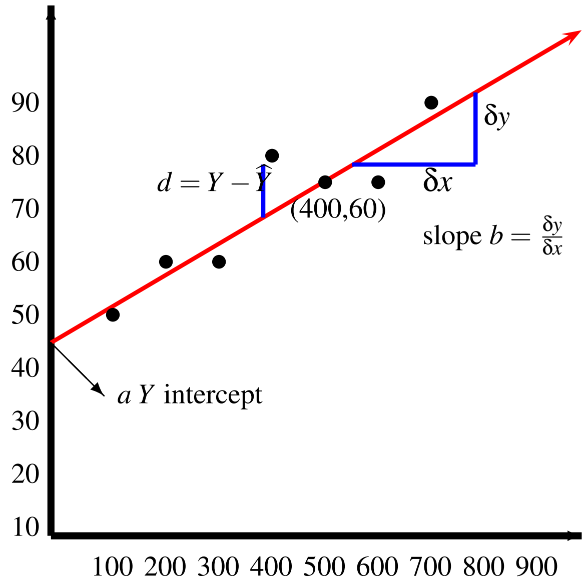
- Prediction of future observations;
- Assessment of the effect of, or relationship between explanatory variables on the response;
- A general description of data structure.

Example

For a particular kind of insurance we want to study of how premiums depend on claims. Let X denote claims and Y denote premiums. A set of seven different levels is shown in the following table:

X	100	200	300	400	500	600	700
Y	40	50	50	70	65	65	80

- Graph these points and roughly fit a line by eye.



Ordinary least squares

The objective is to fit a line whose equation is of the form

$$\widehat{Y} = a + bX.$$

That is, we must find a formula to calculate the slope b and intercept a. This formula derives from the minimization of the sum of squares of all deviations, i.e.

$$\text{minimize} \sum d^2 = \sum (Y - \widehat{Y})^2.$$

This is called the criterion of *Ordinary Least Squares* (OLS) and it selects a unique line called the OLS line.

The OLS slope b is calculated from the formula

$$\boxed{b = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2} = \frac{\sum xy}{\sum x^2},}$$

where $x = X - \bar{X}$, $y = Y - \bar{Y}$ and $x^2 = (X - \bar{X})^2$.

The intercept a can be found from

$$\boxed{a = \bar{Y} - b\bar{X}.}$$

Note that the least-squares line goes through $(\bar{X}, \bar{Y})$.

Example

Using the values of the previous example we have:

$$\bar{X} = 400, \quad \text{and} \quad \bar{Y} = 60.$$

Furthermore, $x = X - \bar{X}$, $y = Y - \bar{Y}$, xy and x^2 are calculated by the table:

x	-300	-200	-100	0	100	200	300
y	-20	-10	-10	10	5	5	20
$xy/1000$	6	2	1	0	0.5	1	6
$x^2/1000$	90	40	10	0	10	40	90

Note that $\sum xy = 16500$ and $\sum x^2 = 280000$. Thus,

$$b = \sum xy / \sum x^2 = 0.059 \quad \text{and} \quad a = \bar{Y} - b\bar{X} = 36.4.$$

Hence, the OLS line is given by:

$$\widehat{Y} = 36.4 + 0.059X.$$

Therefore, if $X = 400$, then the predicted premium $\hat{Y}$ is given by

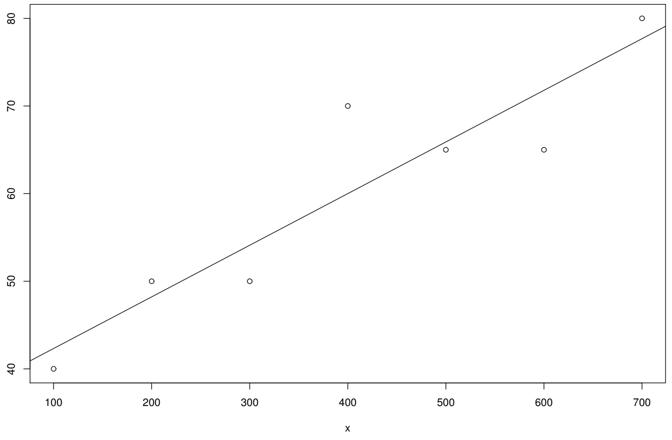
$$\widehat{Y} = 36.4 + 0.059 \times 400 = 60.$$

The deviation d of the actual value Y from the predicted value $\widehat{Y}$ is given by $d = Y - \widehat{Y}$.

Computer fit

```
> x=c(100, 200, 300, 400,500, 600, 700);
> y=c(40, 50, 50, 70,65, 65, 80);
Residuals:
  1     2     3     4     5     6     7
-2.32  1.79 -4.11 10.00 -0.89 -6.79  2.32

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.428     5.038    7.231  0.00079
x             0.059     0.011    5.231  0.00338
---
Residual standard error: 5.961 on 5 df
Multiple R-Squared: 0.85, Adjusted R-squared: 0.81
F-statistic: 27.36 on 1 and 5 DF, p-value: 0.0034
```



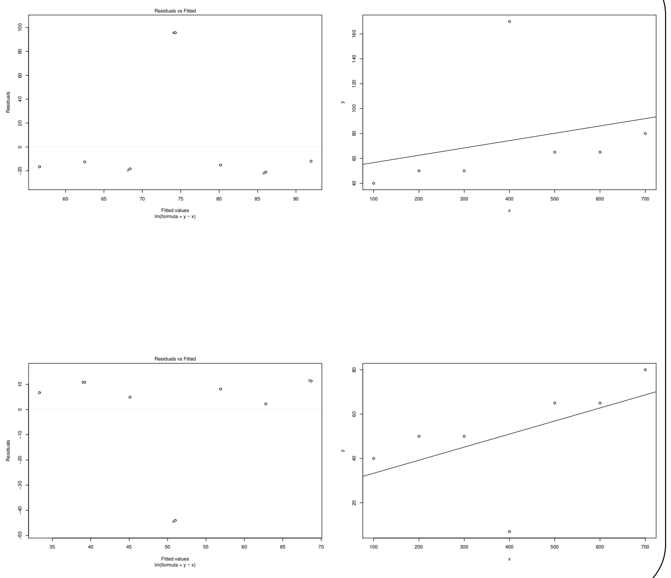
Consider changing the observation $(400,70)$ with $(400,170)$ and $(400,7)$. How the OLS estimators change?

The change: $(400,70)$ to $(400,170)$

```
> y=c(40, 50, 50, 170,65, 65, 80);
Coefficients:
         Estimate Std. Error t value Pr(>|t|)
(Intercept) 50.714   39.19   1.29    0.252
x            0.059    0.09   0.67    0.531
-----
Residual standard error: 46.37 on 5 df
Multiple R-Squared: 0.08, Adjusted R-squared: -0.10
F-statistic: 0.4523 on 1 and 5 DF,  p-value: 0.53
```

The change: $(400,70)$ to $(400,7)$

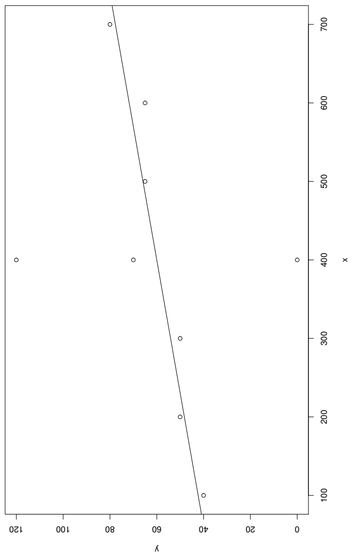
```
> y=c(40, 50, 50, 7,65, 65, 80);
Coefficients:
         Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.429   18.20   1.51    0.192
x            0.059    0.04   1.45    0.207
------
Residual standard error: 21.54 on 5 df
Multiple R-Squared: 0.30, Adjusted R-squared: 0.15
F-statistic: 2.096 on 1 and 5 DF,  p-value: 0.21
```

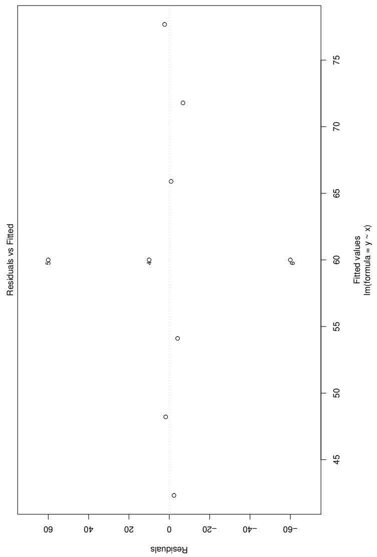


The observation $(400,170)$ and $(400,7)$ are anomalous, but since it occurs near the mean of the explanatory variable, no adverse effects are inflicted on the slope estimate.

Consider adding the new observations $(400, 120)$ and $(400, 0)$.

```
> x=c(100, 200, 300, 400,400,400,500, 600, 700);
> y=c(40, 50, 50, 70,120,0,65, 65, 80);
Coefficients:
       Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.429   26.82   1.36    0.217
x             0.059    0.06   0.96    0.369
--------
Residual standard error: 32.46 on 7 df
Multiple R-Squared: 0.12, Adjusted R-squared: -0.01
F-statistic: 0.92 on 1 and 7 DF,  p-value: 0.37
```





When fitting regression models the following assumptions are made: The (response) random variables $Y_1, \ldots, Y_n$ are independent, with mean $a + bX_i$ and variance σ^2. However, we often write the model in the form

$$Y_i = \alpha + \beta X_i + \varepsilon_i,$$

where ε_i (called error) denotes the deviation of Y from its expected value. In this case the assumptions become: The errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent with mean 0 and variance σ^2.

Sampling variability

We want to investigate how close the estimated line come to the true population line. Particularly, how is the slope estimate b distributed around its target β.

Normal approximation rule for regression

The slope estimate b is approximately normally distributed with mean β and variance $\sigma^2 / \sum x^2$. That is,

$$\boxed{b \sim N(\beta, \sigma^2/\Sigma x^2).}$$

Notice that $\sum x^2 = \sum(X-\bar{X})^2 = nS_x^2$, where S_x^2 is the variance of the variable X. Therefore,

$$\boxed{b \sim N\Big(\beta, \frac{\sigma^2}{nS_x^2}\Big).}$$

The typical deviation of b from its target β represents the estimation error and it is called *Standard Error* (SE). The SE of b is given by

$$\text{SE} = \frac{\sigma}{\sqrt{\Sigma x^2}} = \frac{\sigma}{\sqrt{n}}\,\frac{1}{S_x}.$$

From the latter it follows that there are three ways the SE can be reduced to produce a more accurate estimate b:

1. Reducing σ the inherent variability of the Y observations.
2. increasing the sample size n.
3. Increasing S_x, the spread of the X values which are determined by the experiments (survey).

Consider the true relationship:

$$y = a + bx,$$

where $a = 3.0$ and $b = 5$. The goal is to estimate the relationship when it includes some noise ε.

Suppose that we generate randomly some values for x and a noise ε which is normally distributed with mean zero and standard deviation σ. The y is generated by

$$y = a + bx + \varepsilon.$$

In the R statistical package this is done by:

```
n <- 100
x <- runif(n, min=-200, max=200)
a <- 3.0
b <- 5
e <- rnorm(n, sd=1.5)
y <- a + b * x  + e
g <- lm(y~x)
summary(g)
```

How the estimators are affected by the noise ε and values of x.

1. Let $\sigma = 1.0$ and the range of the x values to be randomly selected from the range -5 to 5. If the sample size is 10, then the OLS estimators of a and b are found to be $\hat{a} = 3.21$ and $\hat{\beta} = 4.78$. If the sample increase to $n = 10000$, then the estimators are $\hat{a} = 2.99$ and $\hat{\beta} = 5.00$.

2. Consider the case where the values of x are randomly selected from the range 5 to 5.2 and $\sigma = 1.0$. With the sample size $n = 10$ the estimators are $\hat{a} = -43.36$ and $\hat{\beta} = 14.00$. For $n = 10000$ it is found that $\hat{a} = 2.34$ and $\hat{\beta} = 5.13$.

3. In the latter case if $\sigma = 0.01$ then for $n = 10$ $\hat{a} = 3.05$ and $\hat{\beta} = 4.99$, while for $n = 10000$ $\hat{a} = 2.99$ and $\hat{\beta} = 5.00$.

4. Increasing the range of the x values to the region -200 to 200 and $\sigma = 1.5$ it gives $\hat{a} = 3.53$ and $\hat{\beta} = 4.99$ when $n = 10$. For $n = 100$ the estimators are found to be $\hat{a} = 2.98$ and $\hat{\beta} = 5.00$

The variance of the Y observations σ^2 is generally unknown and must be estimated. The residuals are used to derive the estimator S^2 of σ^2. That is,

$$\boxed{S^2 = \frac{1}{n-2} \sum (Y - \widehat{Y})^2.}$$

Note that $(n-2)$ is the degrees of freedom and $\sum(Y - \widehat{Y})$ is termed the *sum of squares of errors* (SSE). Thus, $S^2 = \text{SSE}/n-2$, which is an unbiased estimator of σ^2. Therefore, the estimated variance of the slope b is given by $S^2/\sum x^2$.

Furthermore, the 95% confidence interval for β is given by:

$$\boxed{\beta = b \pm T_{2.5\%}^{(n-2)} \frac{S}{\sqrt{\Sigma_{x^2}}}.}$$

Example

From the previous example we have $\widehat{Y} = 36.4 + 0.059 \times 400 = 60$. Hence:

$\widehat{Y}$	42.3	48.2	54.1	60.0	65.9	71.8	77.7
$Y - \widehat{Y}$	-2.3	1.8	-4.1	10.0	-0.9	-6.8	2.3
$(Y - \widehat{Y})^2$	5.29	3.24	16.81	100	0.81	46.24	5.29

Note that SSE $= \sum(Y - \widehat{Y})^2 = 177.68$ and

$$S^2 = \frac{SSE}{n-2} = \frac{177.68}{5} = 35.5$$

$$\frac{S}{\sqrt{\Sigma x^2}} = \sqrt{\frac{35.5}{280000}} = 0.0113$$

$$T_{2.5\%}^{(n-2)} = T_{2.5\%}^{(5)} = 2.571.$$

The latter gives:

$$\begin{aligned} \beta &= b \pm T_{2.5\%}^{(n-2)} \frac{S}{\sqrt{\Sigma x^2}}. \\ &= 0.059 \pm 2.571 \times 0.0113 \\ &= 0.059 \pm 0.29, \end{aligned}$$

or

$$0.030 < \beta < 0.088.$$

The hypothesis that X (claims) and Y (premiums) are unrelated may be stated mathematically as $\beta = 0$. However, at 5% error level we note that zero is not contained in the 95% confidence interval.

Therefore, *at 5% error level we reject the hypothesis that premiums are unrelated to claims.*

P-value

Each statistical test has an associated null hypothesis, denoted by H_0. Null Hypothesis are typically statements of no difference or effect. The p-value is the probability that the sample could have been drawn from the population being tested (or that a more improbable sample could be drawn) given the assumption that the null hypothesis is true. A p-value of 0.05, for example, indicates that you would have only a 5% chance of drawing the sample being tested if the null hypothesis was actually true.

A p-value close to zero signals that the null hypothesis is false, and typically that a difference is very likely to exist. Large p-values closer to 1 imply that there is no detectable difference for the sample size used. A p-value of 0.05 is a typical threshold used in industry to evaluate the null hypothesis. In more critical industries (health-care, etc.) a more stringent, lower p-value may be applied.

To calculate a p-value, collect sample data and calculate the appropriate test statistic for the test you are performing. For example, t-statistic for testing means, Chi-Square or F-statistic for testing variances etc. Using the theoretical distribution of the test statistic, find the area under the curve (for continuous variables) in the direction(s) of the alternative hypothesis (H_1) using a look up table.

Example

What is the p-value for the null hypothesis that premiums DO NOT increase with claims.

Under the null hypothesis we calculate the t-statistic:

$$t = \frac{b}{\text{SE}} = \frac{0.059}{0.0113} = 5.2\,.$$

From tables it can be observed that for the degrees of freedom 5 the t value of 5.2 lies beyond $T_{2.5\%} = 4.77$. Thus,

$$\text{p-value} < 0.0025.$$

This provides so little credibility for H_0 that we could reject it and conclude that premiums do indeed increase with claims.

Note that the alternative hypothesis to the example above is that premiums do increase with claims, That is,

$$H_1 : \quad \beta > 0.$$

Consider the null hypothesis that *premiums are unrelated to claims* (i.e. Y is unrelated to X). This implies that the alternative hypothesis that premiums are related to claims either a positive or a negative way. Thus, we may write this alternative hypothesis as:

$$H_1 : \quad \beta > 0 \text{ or } \beta < 0,$$

or equivalently

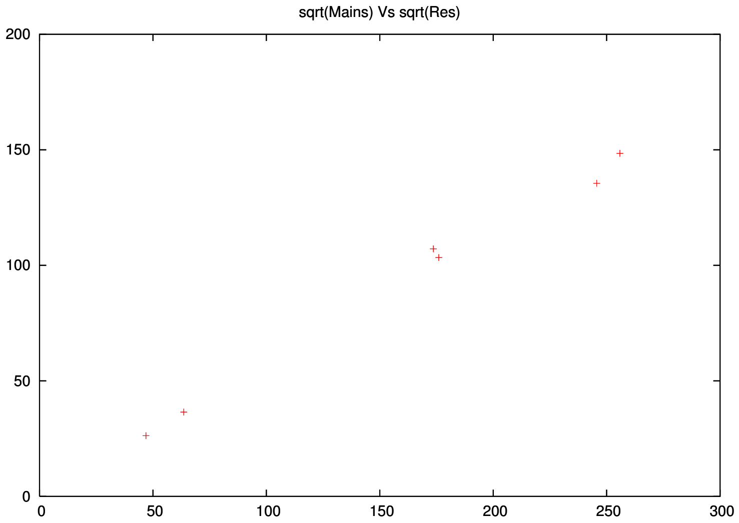
$$H_1 : \quad \beta \neq 0.$$

This is a two-sided hypothesis and thus a two-sided p-value needs to be calculated.

Special Case

The table below shows the population of zones (*Res.*) and the numbers of household mains (*Mains*). We wish to find the relationship of how population size affects the number of telephones. Models connecting these two variables have been used to estimate population in small areas for non-census years.

# Res.	4041	2200	30148	60324	65468	30988
# Mains	1332	690	11476	18368	22044	10686
# $\sqrt{\text{Res.}}$	63.57	46.90	173.63	245.60	255.87	176.03
# $\sqrt{\text{Mains}}$	36.50	26.27	107.13	135.53	148.47	103.37



Let $y = \sqrt{\text{\# of telephones}}$ and $x = \sqrt{\text{population size}}$. The plot indicates a linear relationship with the line passing through $(0,0)$[a].

Consider the regression:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where $\varepsilon_i \sim N(0,\sigma^2)$, and thus, $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$.

The least-squares estimates of β_0 and β_1 are denoted respectively by $b_0 = \bar{y} - b_1\bar{x}$ and $b_1 = \sum xy / \sum x^2$. The b_i is a linear combinations of y_i's and is also normal.

Let the SE (standard error) of b_j denoted by $\text{SE}(b_j)$. Then

$$\frac{(b_j - \beta_j)}{\text{SE}(b_j)} \sim \text{T}^{(n-2)}.$$

Furthermore, the $(1-a) \times 100$ percent C.I. for β_j is given by

$$b_j - \text{SE}(b_j)\,\text{T}_{\frac{a}{2}}^{(n-2)} < \beta_j < b_j + \text{SE}(b_j)\,\text{T}_{\frac{a}{2}}^{(n-2)},$$

where $j = 0,1$ and $\text{T}_{\frac{a}{2}}^{(n-2)}$ is the upper $a/2$ point of the t-distribution with $n-2$ degrees of freedom.

[a] It is perfectly reasonable since if there were no people in an area, then would usually be no household phones.

Computer output:

Variable	j	b_j	$\mathrm{SE}(b_j)$	$t(b_j)$	$P(\lvert t\rvert > \lvert t(b_j)\rvert)$
Intercept	0	1.301	4.280	0.3037	0.7763
$\sqrt{\text{Mains}}$	1	0.571	0.024	23.955	0.0001

That is,

$$b_0 = 1.301,\ b_1 = 0.571,\ \mathrm{SE}(b_0) = 4.28 \text{ and } \mathrm{SE}(b_1) = 0.024.$$

Also,

$$t(b_0) = \frac{b_0}{\mathrm{SE}(b_0)} = 0.3037 \quad \text{and} \quad t(b_1) = \frac{b_1}{\mathrm{SE}(b_1)} = 23.955.$$

Since the $T_{5\%}^{(4)} = 2.1318$, the 90% confidence intervals for β_0 and β_1 are given, respectively, by:

$$(-7.8241, 10.4241) \qquad \text{and} \qquad (0.5198, 0.6221).$$

The interval of β_0 is under the assumption that β_1 is fixed and vice-versa.

Since 0 is included in the interval of β_0 we cannot reject the $H_0 : \beta_0 = 0$. However, we can reject $H_0 : \beta_1 = 0.7$.

The probability that the value of a t-distributed random variable would be numerically larger than $|t(b_0)| = 0.3037$ is 0.7763 and that of getting a t-value larger than $|t(b_1)| = 23.995$ is 0.0001. Thus, we can reject $H_0 : \beta_1 = 0$ at 5, 1, or 0.1 per cent. However, we cannot reject $H_0 : \beta_0 = 0$ at any reasonable level of significance.

When the intercept (β_0) is missing the the computer output is given by:

Variable	j	b_j	$\mathrm{SE}(b_j)$	$t(b_j)$	$P(\lvert t\rvert > \lvert t(b_j)\rvert)$
$\sqrt{\text{Mains}}$	1	0.578	0.0097	59.566	0.0001

The $T_{5\%}^{(5)} = 2.0151$ and thus, the 90% C.I. for β_1 is given by:

$$(0.5583, 0.5973).$$

Goodness of fit

The coefficient of determination R^2 (referred to as *R Squared*) is a measure of *goodness of fit* of the regression line.

Consider the following terminology:

- Total Sum of Squares (TSS): $\sum(Y-\bar{Y})^2$.
- Regression Sum of Squares (RSS): $\sum(\widehat{Y}-\bar{Y})^2$.
- Sum of squares of errors (SSE): $\sum(Y-\widehat{Y})^2$.

We have:

$$\mathrm{TSS} = \mathrm{RSS} + \mathrm{SSE},$$

$$\boxed{R^2 = \frac{\mathrm{RSS}}{\mathrm{TSS}} = \frac{\mathrm{RSS}}{\mathrm{RSS + SSE}} = 1 - \frac{\mathrm{SSE}}{\mathrm{TSS}}}$$

and

$$\boxed{0 \le R^2 \le 1.}$$

A value of $R^2 = 1$ indicates that all the sample observations lie exactly on the regression line, while $R^2 = 0$ indicates that the regression line is of no use at all. I.e. X does not influence Y (linearly) at all.

Example

Using the *Claims-Premiums* example TSS $=1150$, SSE $=177.68$, and thus

$$R^2 = \frac{1150-177.68}{1150} = 0.845.$$

This is interpreted as 84.5% of the variation in premiums (Y) being explained by variation in claims (X). This is quite a respectable figure to obtain, leaving only 15.5% of the variation in premiums left to be explained by other factors.

Note (this should have been said earlier):

The least-squares estimator of β_0 in the simple regression $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ is given by $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$. Thus the variance of $\hat{\beta}_0$ is given by:

$$\begin{aligned}\mathrm{Var}(\beta_0) &= \mathrm{Var}(\bar{y}-\hat{\beta}_1\bar{x}) = \mathrm{Var}(\bar{y}) + \bar{x}^2\mathrm{Var}(\hat{\beta}_1)\\ &= \frac{1}{n^2}\sum \mathrm{Var}(y_i) + \bar{x}^2\frac{\sigma^2}{\sum x_i^2} = \frac{n\sigma^2}{n^2} + \bar{x}^2\frac{\sigma^2}{\sum x_i^2}\\ &= \sigma^2\Big(\frac{1}{n} + \frac{\bar{x}^2}{\sum x_i^2}\Big).\end{aligned}$$

Thus, $\quad \mathrm{SE}(\hat{\beta}_0) = S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum x_i^2}}.$

Overall Significance test

Consider the regression:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

To see if there is any linear relationship we test:

$$\begin{aligned} H_0 : &\quad \beta_0 = \beta_1 = 0 \\ H_1 : &\quad \beta_0 \neq 0 \quad \text{and} \quad \beta_1 \neq 0 \end{aligned}$$

For this test we compute the F-statistic:

$$F = \frac{\text{RSS}}{\text{SSE}/(n-2)}$$

Reject H_0 when F exceeds $F_{\alpha\%}^{(1,n-2)}$, where $(1, n-2)$ are the degrees of freedom of the F distribution and α is the selected percentage point.

In the *claims and premiums* example the F-statistic is computed by

$$\frac{\text{RSS}}{\text{SSE}/(n-2)} = \frac{972.32}{177.67/5} = 27.36.$$

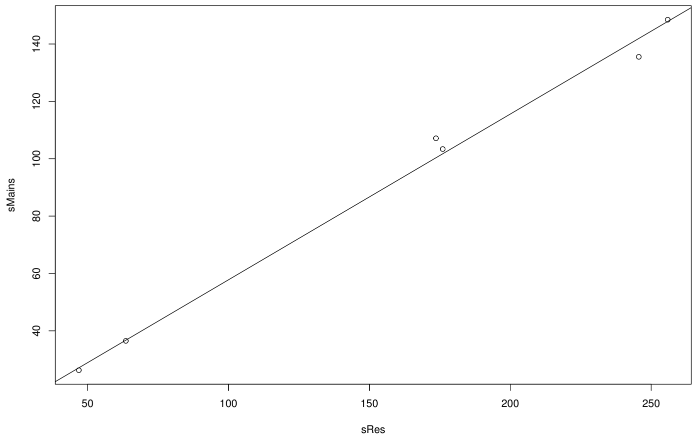
The $F_{2.5\%}^{(1,5)} = 10.01$. Thus, the H_0 can be rejected.

Computer fit (Mains and Residence)

```
Res=c(4041, 2200, 30148, 60324, 65468, 30988)
Mains=c(1332, 690, 11476, 18368, 22044, 10686)
sRes=sqrt(Res)      sMains=sqrt(Mains)

> reg <- lm(sMains~sRes)
sRes =   63.57  46.90 173.63 245.61 255.87 176.03
sMains = 36.50  26.27 107.13 135.53 148.47 103.37
Residuals: 1      2     3     4     5     6
         -1.13 -1.83  6.61 -6.11  0.97  1.49
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.30    4.28   0.30    0.776
sRes         0.57    0.02  23.96  1.8e-05 ***
---
Residual standard error: 4.71 on 4 DF
Multiple R-Squared: 0.99, Adjusted R-squared: 0.99
F-statistic: 573.8 on 1 and 4 DF,  p-value: 1.8e-05

> reg <- lm(sMains~sRes-1)
Residuals: 1      2     3    4     5     6
         -0.24 -0.84  6.79 -6.40  0.61  1.65
Coefficients:
     Estimate Std. Error t value Pr(>|t|)
sRes 0.58   0.010   59.56 2.53e-08 ***
---
Residual standard error: 4.26 on 5 DF
Multiple R-Squared: 1.0, Adjusted R-squared: 0.99
F-statistic:  3547 on 1 and 5 DF,  p-value: 2.5e-08
```



Computer fit (Claims and Premiums)

```
> x=c(100, 200, 300, 400,500, 600, 700);
> y=c(40, 50, 50, 70,65, 65, 80);
Residuals:
  1     2     3     4     5     6     7
-2.32  1.79 -4.11 10.00 -0.89 -6.79  2.32

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.428     5.038    7.231   0.00079
x            0.059     0.011    5.231   0.00338
---
Residual standard error: 5.961 on 5 df
Multiple R-Squared: 0.85, Adjusted R-squared: 0.81
F-statistic: 27.36 on 1 and 5 DF, p-value: 0.003
```

Multiple Regression

Source: Long-Kogan Realty, Chicago, USA.

y	PRICE	Selling price of house in thousands of dollars
X_1	BDR	Number of bedrooms
X_2	FLR	Floor space in sq.ft.
X_3	FP	Number of fireplaces
X_4	RMS	Number of rooms
X_5	ST	Storm windows (1 if present, 0 if absent)
X_6	LOT	Front footage of lot in feet
X_7	TAX	Annual taxes
X_8	BTH	Number of bathrooms
X_9	CON	Construction (0 if frame, 1 if brick)
X_{10}	GAR	Garage size (0 = no garage, 1 = one-car garage, etc.)
X_{11}	CDN	Condition (1 = 'need work', 0 otherwise)
X_{12}	L1	Location (L1 = 1 if property is in zone A, L1 = 0 otherwi
X_{13}	L2	Location (L2 = 1 if property is in zone B, L2 = 0 otherwi

$\text{Price} = f(\text{FLR, ST, LOT, CON, GAR, L2})$

```
26 x 13 (26 observations and 13 exogenous variables)

Y   x1  x2   x3 x4 x5   x6   x7   x8  x9 x10 x11 x12 x13
-------------------------------------------------------
53  2   967  0  5  0  39   652  1.5  1  0.0  0  1  0
55  2   815  1  5  0  33  1000  1.0  1  2.0  1  1  0
56  3   900  0  5  1  35   897  1.5  1  1.0  0  1  0
58  3  1007  0  6  1  24   964  1.5  0  2.0  0  1  0
64  3  1100  1  7  0  50  1099  1.5  1  1.5  0  1  0
44  4   897  0  7  0  25   960  2.0  0  1.0  0  1  0
49  5  1400  0  8  0  30   678  1.0  0  1.0  1  1  0
70  3  2261  0  6  0  29  2700  1.0  0  2.0  0  1  0
72  4  1290  0  8  1  33   800  1.5  1  1.5  0  1  0
82  4  2104  0  9  0  40  1038  2.5  1  1.0  1  1  0
85  8  2240  1 12  1  50  1200  3.0  0  2.0  0  1  0
45  2   641  0  5  0  25   860  1.0  0  0.0  0  0  1
47  3   862  0  6  0  25   600  1.0  1  0.0  0  0  1
49  4  1043  0  7  0  30   676  1.5  0  0.0  0  0  1
56  4  1325  0  8  0  50  1287  1.5  0  0.0  0  0  1
60  2   782  0  5  1  25   834  1.0  0  0.0  0  0  1
62  3  1126  0  7  1  30   734  2.0  1  0.0  1  0  1
64  4  1226  0  8  0  37   551  2.0  0  2.0  0  0  1
66  2   929  0  5  0  30  1355  1.0  1  1.0  0  0  1
35  4  1137  1  7  0  25   561  1.5  0  0.0  0  0  0
38  3   743  0  6  0  25   489  1.0  1  0.0  0  0  0
43  3   596  0  5  0  50   752  1.0  0  0.0  0  0  0
46  2   803  0  5  0  27   774  1.0  1  0.0  1  0  0
46  2   696  0  4  0  30   440  2.0  1  1.0  0  0  0
50  2   691  0  6  0  30   549  1.0  0  2.0  1  0  0
65  3  1023  0  7  1  30   900  2.0  1  1.0  0  1  0
```

Regression model in matrix form

Assume that we have n exogenous variables and m observations. The regression model can be written as:

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_n x_{1n} + \varepsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_n x_{2n} + \varepsilon_2 \\
\vdots\ & \quad \vdots \qquad\qquad\qquad\qquad\qquad\qquad \vdots \\
y_m &= \beta_0 + \beta_1 x_{m1} + \beta_2 x_{m2} + \cdots + \beta_n x_{mn} + \varepsilon_m
\end{aligned}$$

The ith observation can be written as:

$$y_i = \begin{pmatrix} 1 & x_{i1} & x_{i2} & \cdots & x_{in} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} + \varepsilon_i$$

and the whole system of observations can be written as:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{pmatrix},$$

or

$$y = X\beta + \varepsilon.$$

Example (Claims and Premiums)

Consider the simple regression

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2)$$

where X denote claims, Y denote premiums and

x	100	200	300	400	500	600	700
y	40	50	50	70	65	65	80

The regression in matrix form can be written as:

$$\begin{pmatrix} 40 \\ 50 \\ 50 \\ 70 \\ 65 \\ 65 \\ 80 \end{pmatrix} = \begin{pmatrix} 1 & 100 \\ 1 & 200 \\ 1 & 300 \\ 1 & 400 \\ 1 & 500 \\ 1 & 600 \\ 1 & 700 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \end{pmatrix}$$

The latter is equivalent to

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_7)$$

where

$$y = \begin{pmatrix} 40 \\ 50 \\ 50 \\ 70 \\ 65 \\ 65 \\ 80 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 100 \\ 1 & 200 \\ 1 & 300 \\ 1 & 400 \\ 1 & 500 \\ 1 & 600 \\ 1 & 700 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \text{and} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \end{pmatrix}.$$

Ordinary least squares (OLS) estimates

- Consider the linear multiple regression model:

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

  where $y \in \Re^m$, $X \in \Re^{m \times n}$, $\beta \in \Re^n$ and $\varepsilon \in \Re^m$.

- The most frequently used estimating technique for the model (1) is least squares.
- The least squares estimator of β is obtained from solving the normal equations:

$$\boxed{(X^T X)\widehat{\beta} = X^T y.}$$

- The matrix $(X^T X)$ has dimension $n \times n$. It has an inverse if all the exogenous variables are linearly independent, that is, X is of full rank.
- Premultiplying each side of the normal equations by $(X^T X)^{-1}$ it gives

$$(X^T X)^{-1}(X^T X)\widehat{\beta} = (X^T X)^{-1} X^T y,$$

  or the equivalent expression:

$$\boxed{\widehat{\beta} = (X^T X)^{-1} X^T y.}$$

- The OLS estimator $\widehat{\beta}$ is unique.

Examples

In the example of *Claims and Premiums* the vector y and matrix X are given by:

$$y^T = (40\ \ 50\ \ 50\ \ 70\ \ 65\ \ 65\ \ 80) \quad \text{and}$$

$$X^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 100 & 200 & 300 & 400 & 500 & 600 & 700 \end{pmatrix}.$$

Thus,

$$X^T X = \begin{pmatrix} 7 & 2800 \\ 2800 & 1400000 \end{pmatrix}, \qquad X^T y = \begin{pmatrix} 420 \\ 184500 \end{pmatrix}$$

$$(X^T X)^{-1} = \frac{1}{196} \begin{pmatrix} 140 & -2.8 \\ -2.8 & 7 \times 10^{-4} \end{pmatrix} \quad \text{and}$$

$$\hat{\beta} = (X^T X)^{-1} X^T y = \begin{pmatrix} 36.43 \\ 0.059 \end{pmatrix}.$$

Generally, if $n = 2$, then:

$$X^T X = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}.$$

Notice that the condition number of X is given by $\text{Cond}(X) = 1000.0$. If the variable x (claims) is divided by 100, then the condition number becomes 10.404.

Example

Several packages are avaliable to for computing the least-squares and other quantities of interests (SPSS, SAS, GLIM, S-PLUS, R, EXCEL, etc.). For the *House prices* data set the regression equation (not all the variables have been used) is given by:

$$\begin{aligned} \text{PRICE} &= 18.48 + 0.18\,\text{FLR} + 4.03\,\text{RMS} - 7.75\,\text{BDR} \\ &\quad + 2.20\,\text{BTH} + 1.37\,\text{GAR} + 0.257\,\text{LOT} + 7.09\,\text{FP} + 10.96\,\text{ST}. \end{aligned}$$

Consider the estimated selling price of a house with 1000 square feet of floor area, 8 rooms, 4 bedrooms, 2 baths, storm window, no fireplaces, 40 foot frontage and 1 car garage:

$$\begin{aligned} &18.48 + 0.18(1000) + 4.03(8) - 7.75(4) + 2.20(2) + 1.37(1) \\ &\quad + 0.257(40) + 7.09(0) + 10.96(1) = 64.73. \end{aligned}$$

From the regression it can be observed that:

- An additional car in a garage would raise the price by about \$1370.
- Every square foot increase in floor area increases the price by about \$18.

Each of these price changes is marginal, i.e. nothing else changes.

Observe the negative sign associated with the bedrooms (BDM). This implies an estimated loss of prices occurs if we increase the number of bedrooms without increasing the number of rooms and floor area. E.g. if in addition we increase the number of rooms by one, add a bathroom and some floor area, then the estimated price will go up.

In situations where there are several related variables, signs which at first glance would appear counter-intuitive are not uncommon. A further investigation might show an explanation of this plausibility.

Furthermore, the estimates are random variables and even more importantly, we may not have considered important variables. That is, we are far from the truth model.

It is true that a perfect model is seldom possible.

Assumptions of the standard linear regression model

Consider the regression:

$$y_i = X_i\beta + \varepsilon_i \quad \text{or} \quad y = X\beta + \varepsilon.$$

In order for the estimates of β to have some statistical properties we need to make some assumptions about how the observations y have been generated.

- $\mathrm{E}(\varepsilon) = 0.$ That is, $\mathrm{E}(y) = X\beta.$

Assume that X variables measure family income and various other family characteristics and Y denotes family expenditure on travel. The first row of the X matrix is some specific set of numbers for family income, size and composition. Let s_1 denote a row vector consisting of these numbers. Then the average, or expected level of travel expenditure for this type of family is given by:

$$\mathrm{E}(y_1) = s_1\beta.$$

However, the *actual* travel expenditure of families with these characteristics may be greater, or less that the expected value. Furthermore, in different periods the expenditure of the same family will fluctuate around the mean value. *If all the significant variables are included in X, then we expect that the positive and negative discrepancies from the expected value will occur and they will average to zero*. That is, $\mathrm{E}(\varepsilon_1) = 0$.

Similar considerations apply to each row of X, and so we have:

$$\mathrm{E}(\varepsilon) = \begin{pmatrix} \mathrm{E}(\varepsilon_1) \\ \mathrm{E}(\varepsilon_2) \\ \vdots \\ \mathrm{E}(\varepsilon_m) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = 0.$$

- $\mathrm{E}(\varepsilon\varepsilon^T) = \sigma^2 I_m$.

The variance matrix of ε is given by
$\mathrm{E}\Big((\varepsilon - \mathrm{E}(\varepsilon))((\varepsilon - \mathrm{E}(\varepsilon))^T\Big) = \mathrm{E}(\varepsilon\varepsilon^T) \quad$ since $\mathrm{E}(\varepsilon) = 0$.

Thus,

$$\mathrm{E}(\varepsilon\varepsilon^T) = \begin{pmatrix} \mathrm{Var}(\varepsilon_1) & \mathrm{Cov}(\varepsilon_1,\varepsilon_2) & \dots & \mathrm{Cov}(\varepsilon_1,\varepsilon_m) \\ \mathrm{Cov}(\varepsilon_2,\varepsilon_1) & \mathrm{Var}(\varepsilon_2) & \dots & \mathrm{Cov}(\varepsilon_2,\varepsilon_m) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(\varepsilon_m,\varepsilon_1) & \mathrm{Cov}(\varepsilon_m,\varepsilon_2) & \dots & \mathrm{Var}(\varepsilon_m) \end{pmatrix}$$

$$= \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix} = \sigma^2 I_m$$

This is a double assumption, namely:

1. Each ε_i distribution has the same variance.

   This property is referred as *homoscedasticity* (homogeneous variances) and its opposite as *heteroscedasticity*.

   E.g. If we consider a cross section of the population, then the assumption of *heteroscedasticity* might be more reasonable. This is because low income families will almost certainly have low average expenditures on travel and also low variance of actual travel expenditure about the average. On the other hand high income families will tend to display both higher mean levels of expenditure and greater variance about the mean.

2. All disturbances are pairwise uncorrelated.

   This is a strong assumption. This assumption implies for example that high expenditure in one year does not tend to be associated with usually low (or high) expenditure in the next year, or subsequent years. Another example, is that the assumption denies the possibility of *keeping up with the neighbor*. That is, the size of the disturbance of one family does not have an influence on the size of the disturbance for another family.

- The X is a nonstochastic matrix: $\mathrm{E}(X^T\varepsilon) = 0$.

  This means that if we take another sample of n observations, then the X matrix of explanatory variables remains unchanged. The only source of variation then being in ε and hence in y.

Mean and variance of estimates

Consider the regression:

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_m).$$

The OLS estimator is given by:

$$\widehat{\beta} = (X^TX)^{-1}X^Ty.$$

Substituting $y = X\beta + \varepsilon$ in the latter it gives:

$$\begin{aligned}\widehat{\beta} &= (X^TX)^{-1}X^Ty \\ &= (X^TX)^{-1}X^T(X\beta + \varepsilon) \\ &= (X^TX)^{-1}X^TX\beta + (X^TX)^{-1}X^T\varepsilon \\ &= \beta + (X^TX)^{-1}X^T\varepsilon.\end{aligned}$$

Thus,

$$\begin{aligned}\mathrm{E}(\widehat{\beta}) &= \mathrm{E}(\beta + (X^TX)^{-1}X^T\varepsilon) \\ &= \mathrm{E}(\beta) + (X^TX)^{-1}X^T\mathrm{E}(\varepsilon) \\ &= \beta.\end{aligned}$$

Note that $\widehat{\beta} - \mathrm{E}(\widehat{\beta}) = \widehat{\beta} - \beta = (X^TX)^{-1}X^T\varepsilon$.

Thus,

$$\begin{aligned}
\mathrm{Var}(\widehat{\beta}) &= \mathrm{E}\Big((\widehat{\beta} - \mathrm{E}(\widehat{\beta}))(\widehat{\beta} - \mathrm{E}(\widehat{\beta}))^T\Big) \\
&= \mathrm{E}\Big((X^TX)^{-1}X^T\varepsilon\varepsilon^TX(X^TX)^{-1}\Big) \\
&= (X^TX)^{-1}X^T\mathrm{E}(\varepsilon\varepsilon^T)X(X^TX)^{-1} \\
&= (X^TX)^{-1}X^T\sigma^2I_mX(X^TX)^{-1} \\
&= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} \\
&= \sigma^2(X^TX)^{-1}.
\end{aligned}$$

The elements in the main diagonal of $\mathrm{Var}(\widehat{\beta}) = \sigma^2(X^TX)^{-1}$ give the sampling variances of the corresponding elements of $\widehat{\beta}$.

Estimation of σ^2

Usually σ^2 is not know and needs to be estimated in order to make various inferences. This can be done using the residuals e_i. An unbiased estimator of σ^2 is given by:

$$\boxed{s^2 = \frac{1}{m-n-1}\sum_{i=1}^{n} e_i^2 = \frac{e^Te}{m-n-1}.}$$

Example (Claims and Premiums)

The estimator of σ^2 is found to be $S^2 = 35.53$. Furthermore,

$$X^TX = \begin{pmatrix} 7 & 2800 \\ 2800 & 1400000 \end{pmatrix}$$

and

$$\begin{aligned}
S^2(X^TX)^{-1} &= 35.53 \times \frac{1}{196}\begin{pmatrix} 140 & -2.8 \\ -2.8 & 7\times 10^{-4} \end{pmatrix} \\
&= \begin{pmatrix} 25.38 & -0.051 \\ -0.051 & 0.0001 \end{pmatrix}.
\end{aligned}$$

The diagonal entries of $S\sqrt{(X^TX)^{-1}}$ are given by:

5.038 and 0.011.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.428    5.038   7.231   0.00079
x            0.059    0.011   5.231   0.00338
---
Residual standard error: 5.961 on 5 df
```

Example (House data)

Consider fitting the model:

$$\text{PRICE}=\beta_1\text{FLR}+\beta_2\text{RMS}+\beta_3\text{BDR}+\beta_4\text{GAR}+\beta_5\text{ST}+\varepsilon.$$

The response variable $y = \text{PRICE}$ and the 26×5 exogenous matrix X are given by:

$$y = \begin{pmatrix} 53 \\ 55 \\ \vdots \\ 65 \end{pmatrix} \quad \text{and} \quad X = (\text{FLR} \;\; \text{RMS} \;\; \text{BDR} \;\; \text{GAR} \;\; \text{ST}) = \begin{pmatrix} 967 & 5 & 2 & 0.0 & 0 \\ 815 & 5 & 2 & 2.0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1023 & 7 & 3 & 1.0 & 1 \end{pmatrix}$$

Now,

$$X^T y = \begin{pmatrix} 1712260 \\ 9801 \\ 4884 \\ 1376 \\ 458 \end{pmatrix},$$

$$X^T X = \begin{pmatrix} 36714794 & 200359.0 & 102510.0 & 28014.0 & 8368.0 \\ 200359 & 1171.0 & 597.0 & 153.5 & 50.0 \\ 102510 & 597.0 & 314.0 & 77.5 & 26.0 \\ 28014 & 153.5 & 77.5 & 35.5 & 7.5 \\ 8368 & 50.0 & 26.0 & 7.5 & 7.0 \end{pmatrix}$$

and

$$(X^T X)^{-1} = \frac{1}{1000}\begin{pmatrix} 0.00 & -0.06 & -0.02 & -0.05 & 0.03 \\ -0.06 & 37.73 & -50.66 & -3.84 & -5.05 \\ -0.02 & -50.66 & 106.04 & 8.69 & -13.67 \\ -0.05 & -3.84 & 8.69 & 72.76 & -17.17 \\ 0.03 & -5.05 & -13.67 & -17.17 & 211.49 \end{pmatrix}$$

$$\text{Thus,} \quad \begin{pmatrix} \hat\beta_1 \\ \hat\beta_2 \\ \hat\beta_3 \\ \hat\beta_4 \\ \hat\beta_5 \end{pmatrix} \equiv \hat\beta = (X^T X)^{-1} X^T y = \begin{pmatrix} 0.015 \\ 11.401 \\ -12.519 \\ 3.040 \\ 9.413 \end{pmatrix}.$$

The residual $e = y - X\hat\beta$ and the estimator of σ^2 is calculated by $S^2 = e^T e/(m-n)$, where $m = 26$ and $n = 5$. That is,

$$S^2 = 62.98, \quad \text{or} \quad S = 7.94.$$

The diagonal entries of $S\sqrt{(X^T X)^{-1}}$ gives the standard errors of $\hat\beta$, i.e.

$$\begin{pmatrix} 0.005 & 1.542 & 2.584 & 2.141 & 3.650 \end{pmatrix}^T.$$

The *computer fit* gives:

```
Coefficients:
   Estimate Std. Error t value Pr(>|t|)
FLR   0.015   0.005    2.78    0.011
RMS  11.401   1.542    7.40    2.8e-07
BDR -12.519   2.584   -4.85    8.7e-05
GAR   3.040   2.141    1.42    0.17
ST    9.413   3.650    2.58    0.02
---
Residual standard error: 7.94 on 21 DF
Multiple R-Squared: 0.98, Adjusted R-squared: 0.98
```

Gauss-Markov theorem

The OLS estimator $\widehat{\beta} = (X^T X)^{-1} X^T y$ is the *Best Linear unbiased estimator* (BLUE). This implies that:

1. $\mathrm{E}(\widehat{\beta}) = \beta$.

   The linearity refers to y (or ε). I.e. each element of $\widehat{\beta}$ is a linear combination of y (or ε).

2. No other linear unbiased estimator can have smaller sampling variances those of the OLS estimator $\widehat{\beta}$.

The Gauss-Markov theorem states that the least-squares estimator of $\hat{\beta}$ is a good choice. However, if the errors are correlated or have unequal variance, there will be better estimators. In some cases non-linear or biased estimates may work better in some sense. Thus, the theorem does not tell one to use least-squares all the time, it just strongly suggests it unless there is some strong reason to do otherwise. E.g.

1. If the errors are correlated or have unequal variance, then generalized least-squares should be used.

2. When the predictors are highly correlated (collinear), then biased estimators such as ridge regression might be preferable.

Goodness of fit

A statistic that is widely used to determine how well a regression fits is the coefficient of determination R^2. The R^2 explains how much of the variability in the y can be explained by the fact that they are related to X, i.e., how close the points are to the line. The Coefficient of Determination $0 \leq R^2 \leq 1$ is provided by all computer packages. It is defined as:

$$\boxed{R^2 = 1 - \frac{\text{Residual Sum of Squares}}{\text{Total Sum of Squares}}.}$$

Often small sample sizes inflate R^2. The R^2 always increases with the addition of a new variable. Specifically, adding a variable to a model can only decrease the RSS and so only increase R^2.

Thus, R^2 by itself is not a good criterion because it would always choose the largest possible model.

Example (Claims and Premiums)

The claims and premiums are given, respectively, by:

(claims) x	100	200	300	400	500	600	700
(premiums) y	40	50	50	70	65	65	80

The *computer fit* of the linear regression

$$y = \beta_0 + \beta_1 x + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

gives:

```
Residuals:
 1       2    3     4     5     6      7
-2.32  1.79 -4.11 10.00 -0.90 -6.79  2.32
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.428    5.04    7.23   0.00079 ***
x            0.059    0.01    5.23   0.00338 **
---
Residual standard error: 5.96 on 5 DF
Multiple R-Squared: 0.85, Adj R-squared: 0.81, cp=2
F-statistic: 27.36 on 1 and 5 DF,  p-value: 0.0033
```

Consider now generating a random variable z from the uniform distribution between min $= 100$ and max $= 1000$. I.e. $z^T = (129.29, 231.47, 770.31, 127.14, 674.62, 217.54, 278.22)$.

The *computer fit* of the linear regression

$$y = \beta_0 + \beta_1 x + \beta_2 z + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

gives:

```
Coefficients:
           Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.765   5.97    6.46   0.003 **
x             0.060   0.01    5.10   0.007 **
z           -0.008   0.01   -0.81   0.464
---
Residual standard error: 6.18 on 4 DF
Multiple R-Squared: 0.87, Adj. R-squared: 0.80, cp=3
F-statistic: 13.06 on 2 and 4 DF,  p-value: 0.0176
```

The *computer fit* of the linear regression

$$y = \beta_0 + \beta_1 z + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

gives:

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept) 61.087   9.944   6.14  0.002 **
z           -0.003   0.023  -0.13  0.899
---
Residual standard error: 15.14 on 5 DF
Multiple R-Squared: 0.004, Adj. R-squared: -0.20, cp=2
F-statistic: 0.018 on 1 and 5 DF,  p-value: 0.90
```

The Adjusted coefficient of determination R^2 takes into account for the number of variables and sample size. It is defined by:

$$R_a^2 = 1 - \frac{\mathrm{RSS}/(m-n-1)}{\mathrm{TSS}/(m-1)}$$
$$= 1-(1-R^2)\frac{(m-1)}{(m-n-1)}.$$

Observe that R_a^2 can decline if a new variable produces too small a reduction in $1-R^2$.

Mallows C_p.

The deletion of an exogenous variable from a model is usually biases the model. Furthermore, a deletion of a variable also decreases the covariance matrix of the estimates. The C_p (having p independent) variables is defined as:

$$C_p = \frac{\mathrm{RSS}_p}{\hat{\sigma}^2} - (m-2p).$$

If $C_p \approx p$, then the model does not lead to much bias.

The House Prices data set – selected models

Branch and Bound - exhaustive search

# of var.	R^2	Adjusted R^2	C_p	Model
0	0.00	-0.04	160.45	const.
1	0.54	0.50	62.56	FLR
2	0.67	0.63	40.04	FLR ST
3	0.76	0.71	26.38	FLR FP ST
4	0.81	0.76	18.94	BDR FLR FP ST
5	0.87	0.82	10.55	BDR FLR FP RMS ST
6	**0.90**	**0.86**	**6.20**	**FLR ST LOT CON GAR L2**
7	0.92	0.88	4.94	BDR FLR ST LOT CON GAR L2
8	0.93	0.89	4.81	BDR FLR RMS ST LOT CON GAR L2
9	0.94	0.89	5.96	BDR FLR FP RMS ST LOT CON GAR L2
10	0.94	0.89	7.51	BDR FLR FP RMS ST LOT BTH CON GAR L2
11	0.94	0.88	9.28	BDR FLR FP RMS ST LOT BTH CON GAR L1 L2
12	0.94	0.87	11.08	BDR FLR FP RMS ST LOT TAX BTH CON GAR L1 L2
13	0.94	0.86	13.00	BDR FLR FP RMS ST LOT TAX BTH CON GAR CDN L1 L2

Regression diagnostics

In the 1970s and 80s, many statisticians developed techniques for assessing multiple regression models. One of the most influential books on the topic was Regression Diagnostics: Identifying Influential Data and Sources of Collinearity by Belsley, Kuh, and Welch. Roy Welch tells of getting interested in regression diagnostics when he was once asked to fit models to some banking data. When he presented his results to his clients, they remarked that the model could not be right because the sign of one of the predictors was different from what they expected. When Welch looked closely at the data, he discovered the sign reversal was due to an outlier in the data. This example motivated him to develop methods to insure it didn't happen again!

- The goal is to identify remarkable observations and unremarkable predictors.

- Problems with observations, i.e. *Outliers* and *Influential observations*.

  1. An observation (or measurement) that is unusually large or small relative to the other values in a data set is called an outlier. Outliers typically are attributable to one of the following causes:

     a. The measurement is observed, recorded, or entered into the computer incorrectly.

     b. The measurements come from a different population.

     c. The measurement is correct, but represents a rare event.

  2. Influential observations refer to observations that have a substantial influence on the fitted regression function (i.e., the estimated regression function is substantially different depending on whether the observations are included or not in the data set).

     In other words, Influential observations pull the regression line towards themselves and deleting these observations changes your statistical analysis markedly.

- Problem with the predictors. I.e.

  1. A predictor may not add much to the model. In this case model selection techniques could be used.

  2. A predictor may be too similar to another predictor (collinearity). Identify these predictors and/or transform the model. E.g. using PCA.

  3. Predictors may have been left out.

The HAT matrix

- Given the ordinary regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

or in compact form:

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_m).$$

The BLUE of β is given by:

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

- The predicted values of y are given by:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_n x_{in}, \quad i = 1, \ldots, m,$$

or in matrix form:

$$\begin{aligned} \hat{y} &= X\hat{\beta} \\ &= X\big((X^T X)^{-1} X^T y\big) \\ &= X(X^T X)^{-1} X^T y \\ &= Hy \end{aligned}$$

where the $m \times m$ matrix $H = X(X^T X)^{-1} X^T$ is called the *hat matrix*.

- The hat matrix is *idempotent*. That is, $H^T = H$ and $H^2 = H$.

- The variance-covariance of $\hat{y}$ has the form:

$$\text{Var}(\hat{y}) = \begin{pmatrix} \text{Var}(\hat{y}_1) & \text{Cov}(\hat{y}_1, \hat{y}_2) & \ldots & \text{Cov}(\hat{y}_1, \hat{y}_m) \\ \text{Cov}(\hat{y}_2, y_1) & \text{Var}(\hat{y}_2) & \ldots & \text{Cov}(\hat{y}_2, \hat{y}_m) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(\hat{y}_m, \hat{y}_1) & \text{Cov}(\hat{y}_m, \hat{y}_2) & \ldots & \text{Var}(\hat{y}_m) \end{pmatrix}$$

- The variance-covariance of $\hat{y}$ is given by:

$$\begin{aligned} \text{Var}(\hat{y}) &= \text{Var}(Hy) \\ &= H\text{Var}(y)H^T \\ &= H\sigma^2 I_m H \quad (\text{since Var}(y) = \sigma^2 I_m) \\ &= \sigma^2 H^2 \qquad (\text{since } H^T = H \text{ and } I_m H = H) \\ &= \sigma^2 H \qquad (\text{since } H^2 = H). \end{aligned}$$

- The diagonal elements of H gives the variances of $\hat{y}_i$ for $i = 1, \ldots, m$. That is,

$$\text{Var}(\hat{y}_i) = \sigma^2 h_{ii}.$$

- Note that

$$
\begin{aligned}
h_{11} + h_{22} + \cdots + h_{mm} &= \operatorname{trace}(H) \\
&= \operatorname{trace}(X(X^T X)^{-1} X^T) \\
&= \operatorname{trace}((X^T X)^{-1} X^T X) \\
&= \operatorname{trace}(I_n) \\
&= n.
\end{aligned}
$$

- The total of all variances of $\hat{y}_i$ is $n\sigma^2$. I.e.

$$\sum_{i=1}^{m} \operatorname{Var}(\hat{y}_i) = \operatorname{trace}(\sigma^2 H) = \sigma^2 \sum_{i=1}^{m} h_{ii} = n\sigma^2.$$

- The diagonal elements of the *hat matrix* h_{ii} are called *leverages*. The leverages are useful in diagnostics.

  Notice that the average value of h_{ii} is n/m. Thus, a *rule of thumb* is that leverages of more than $2n/m$ should be looked at most closely. Large values of h_{ii} are due to extreme values in X.

$$\boxed{\text{An observation is influential if } h_{ii} > \frac{2n}{m}.}$$

Example (House Data)

Consider fitting the model:

$$\text{PRICE}=\beta_0+\beta_1\text{FLR}+\beta_2\text{RMS}+\beta_3\text{BDR}+\beta_4\text{GAR}+\beta_5\text{ST}+\varepsilon.$$

The *computer fit* gives:

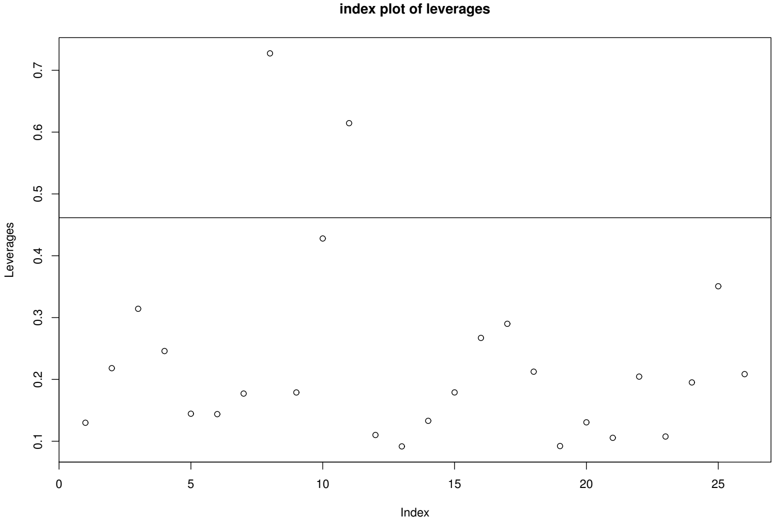
```
Coefficients:
             Estim Std. Error t value Pr(>|t|)
(Intercept) 23.30   5.74   4.06    0.0006 ***
FLR          0.02   0.00   4.15    0.0005 ***
RMS          5.01   1.96   2.55    0.0190 *
BDR         -7.39   2.33  -3.17    0.0049 **
GAR          3.25   1.63   2.00    0.0592 .
ST           9.95   2.77   3.59    0.0018 **
---
Residual standard error: 6.022 on 20 DF
Multiple R-Squared: 0.82, Adjusted R-squared: 0.77
```

The leverages are given by:

```
leverages <- hat(X)
highlev = 2*6/26 = 0.46         sum(leverages) = 6
> leverages
 0.13 0.22 0.31 0.25 0.14 0.14 0.18 0.73 0.18
 0.43 0.61 0.11 0.09 0.13 0.18 0.27 0.29 0.21
 0.11 0.20 0.11 0.20 0.35 0.21 0.09 0.13

> leverages[leverages > highlev]
  8   11
0.73 0.61
```

```
highlev <- 2*6/26                  lowlev <- 1/26.0
plot(leverages, ylab="Leverages",main="index ...)
abline(h=highlev)
```



Deleting the 8th observation it gives:

```
Coefficients:
            Estimate Std. Error t  value Pr(>|t|)
(Intercept) 28.22     5.769    4.89  0.0001 ***
FLR           0.029   0.007    4.33 0.00036 ***
RMS           2.076   2.270    0.92 0.37183
BDR          -6.777   2.169   -3.13 0.00558 **
GAR           3.850   1.523    2.53 0.02053 *
ST            9.239   2.576    3.59 0.00199 **
---
Residual standard error: 5.55 on 19 DF
Multiple R-Squared: 0.84, Adjusted R-squared: 0.80
```

Residuals

- The residuals can also be expressed in terms of the hat matrix:

$$e_i = y_i - \hat{y}_i, \quad i = 1, \ldots, m.$$

In compact form:

$$\begin{aligned} e &= y - \hat{y} \\ &= y - Hy \\ &= (I_m - H)y. \end{aligned}$$

- The residual sum of squares is given by $\sum_{i=1}^{m} e_i^2$, or:

$$\begin{aligned} e^T e &= y^T (I_m - H)^T (I_m - H)y \\ &= y^T (I_m - H)y, \end{aligned}$$

since $(I_m - H)$ is idempotent. That is,
$(I_m - H) = (I_m - H)^T$ and $(I_m - H)^2 = (I_m - H)$.

- The variance-covariance of e has the form:

$$\mathrm{Var}(e) = \begin{pmatrix} \mathrm{Var}(e_1) & \mathrm{Cov}(e_1, e_2) & \ldots & \mathrm{Cov}(e_1, e_m) \\ \mathrm{Cov}(e_2, e_1) & \mathrm{Var}(e_2) & \ldots & \mathrm{Cov}(e_2, e_m) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(e_m, e_1) & \mathrm{Cov}(e_m, e_2) & \ldots & \mathrm{Var}(e_m) \end{pmatrix}$$

- The variance-covariance of e is given by:

$$\begin{aligned}\mathrm{Var}(e) &= \mathrm{Var}\big((I_m - H)y\big) \\ &= (I_m - H)\mathrm{Var}(y)(I_m - H)^T \\ &= (I_m - H)\sigma^2 I_m (I_m - H) \quad (\text{since } \mathrm{Var}(y) = \sigma^2 I_m) \\ &= \sigma^2 (I_m - H)^2 \\ &= \sigma^2 (I_m - H) \qquad (\text{since } (I_m - H)^2 = (I_m - H)).\end{aligned}$$

- The $\mathrm{trace}(I_m - H) = \mathrm{trace}(I_m) - \mathrm{trace}(H) = m - n$.

- The variances of e_i is given by the ith diagonal element of $\sigma^2(I_m - H)$, i.e.

$$\boxed{\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii}), \quad \text{for } i = 1, \ldots, m.}$$

- Notice that $\mathrm{Var}(e_i) \geq 0$ and thus,

$$\boxed{1 - h_{ii} \geq 0, \quad \text{or} \quad h_{ii} \leq 1.}$$

Standardized residuals

- Recall that the variance of the residual $e_i = y_i - \hat{y}_i$ is given by

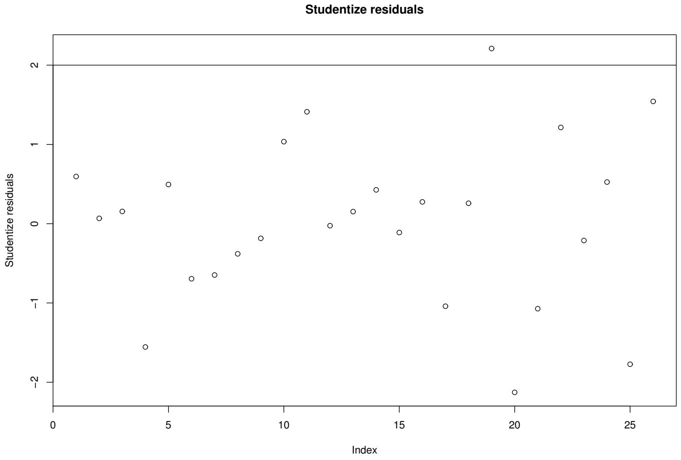
$$\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii}), \quad \text{for } i = 1, \ldots, m.$$

- The (internally) Standardized residuals are given by:

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{(1 - h_{ii})}}.$$

- If the linear regression assumptions are correct, then $\mathrm{Var}(r_i) = 1$ and $\mathrm{Cor}(r_i, r_j)$ tends to be small.

- Outlier if $\|r_i\| > 2$.



The 26th observation of the House data is given by:

```
Price BDR  FLR FP RMS ST LOT  TAX BTH CON GAR CDN L1 L2
   65   3 1023  0   7  1  30  900 2.0   1 1.0   0  1  0
```

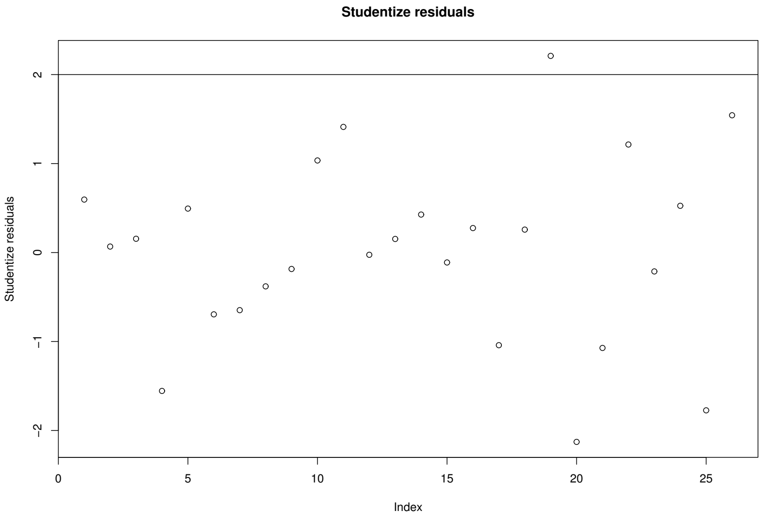
suppose an error occur and the last observation of the House data has been reported as:

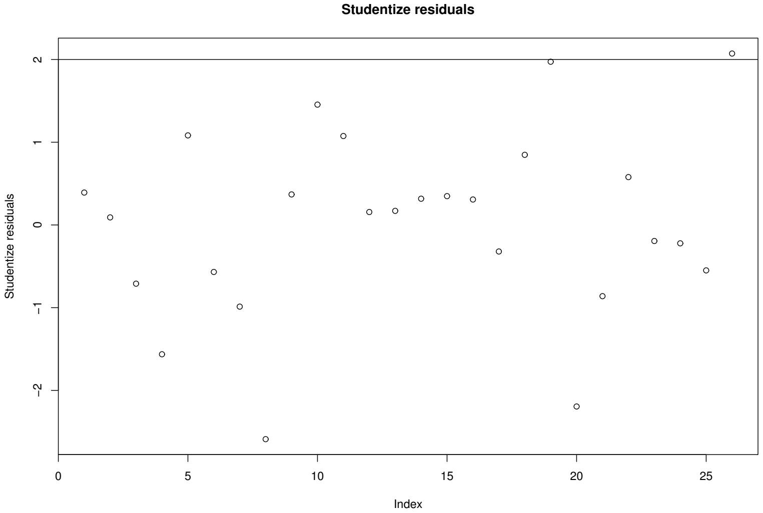
```
Price BDR  FLR FP RMS ST LOT  TAX BTH CON GAR CDN L1 L2
   65   3 1023  0   1  1  30  900 2.0   1 1.0   0  1  0
```

That is, the RMS was replaced by 1 (instead of 7).

The new estimators are given by:

```
Coefficients:
            Estimate Std.  Error t value Pr(>|t|)
(Intercept) 32.69   4.57   7.16 6.21e-07 ***
FLR          0.02   0.01   4.56 0.000190 ***
RMS          1.18   1.21   0.97 0.342454
BDR         -3.70   1.96  -1.89 0.073696 .
GAR          3.49   1.83   1.91 0.070673 .
ST          11.43   3.27   3.50 0.002263 **
---
Residual standard error: 6.78 on 20 DF
Multiple R-Squared: 0.77, Adjusted R-squared: 0.71
```





Influential Observations: Cook's distance

- An influential point is one whose removal from the data set would cause a large change in the fit. An influential point may or may not be an outlier and may or may not have a large leverage, but it will tend to have at least one of those properties.

- Let the subscript i indicates the fit without the observation (i). Here are some measures of influence:

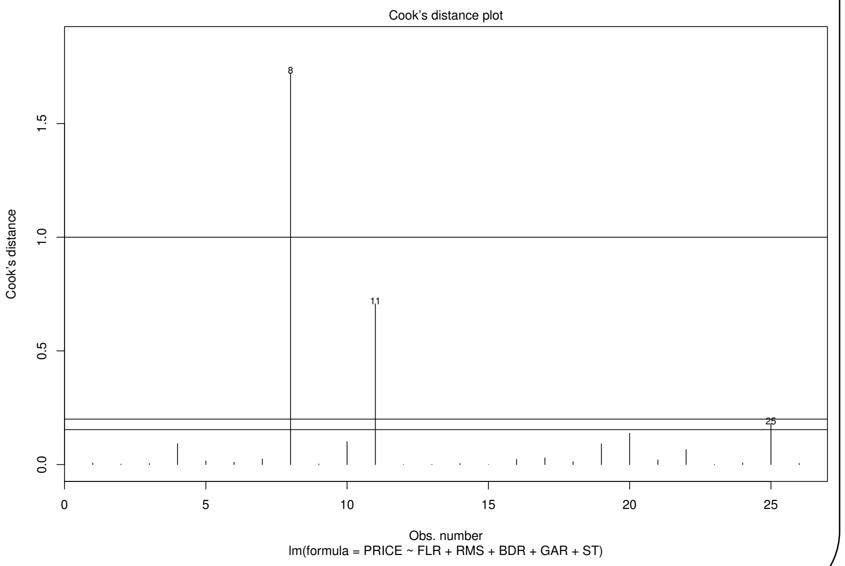
  1. Change in the coefficients: $\hat{\beta} - \hat{\beta}_{(i)}$.

  2. Change in the fit: $\hat{y} - \hat{y}_{(i)} = X^T(\hat{\beta} - \hat{\beta}_{(i)})$.

- These are hard to judge in the sense that the scale varies between datasets. A popular alternative is the Cook's distance:

$$D_i = \frac{(\hat{y} - \hat{y}_{(i)})^T(\hat{y} - \hat{y}_{(i)})}{n\hat{\sigma}^2}$$
$$= \frac{r_i^2}{n}\,\frac{h_{ii}}{1-h_{ii}}.$$

- Cook's distance D_i, is another measure of the influence of a case. Cook's distance measures the effect of deleting a given observation. Observations with larger D_i values than the rest of the data are those which have unusual leverage. A suggested cut-off for detecting influential cases, values of D_i greater than $4/(m-n)$, where m is the number of observations and n is the number of independent variables (including the constant). Others suggest $D_i > 1$ as the criterion to constitute a strong indication of an outlier problem, with $D_i > 4/m$ the criterion to indicate a possible problem.



Collinearity

- The degree to which the independent variables are correlated, and thus predict one another, is collinearity. If collinearity is so high that some of the independent variables almost totally predict other independent variables then this is known as multicollinearity.
- Multicollinearity causes problems in using regression models to draw conclusions about the relationships between predictors and outcome. An individual predictor's p-value may test non-significant even though it is important. Confidence intervals for regression coefficients in a multicollinear model may be so high that tiny changes in individual observations have a large effect on the coefficients, sometimes reversing their signs.
- One obvious method of assessing the degree to which each independent variable is related to all other variables is to examine R_j^2, which is the value of the coefficient of determination R^2 between the variable x_j and all other independent variables. That is, R_j^2 is the R^2 we would get if we regress x_j against all other x_i's.

- The tolerance TOL_j is defined as:

$$\text{TOL}_j = 1 - R_j^2.$$

- TOL_j is closed to 1 if x_j is not closely related to other predictors.
- The Variance Inflation Factor (and the reciprocal, tolerance) as a measure of collinearity:

$$\text{VIF}_i = \frac{1}{1 - R_i^2}.$$

- A value of VIF_i close to 1 indicates no relationship, while larger values indicate presence of multicollinearity (redundant information in the explanatory variables).

  E.g. if $R_j^2 = 0.90$, then $\text{VIF}_i = 10$ and caution is advised (some others say $\text{VIF}_i = 5$, i.e. $R_j^2 = 0.80$).
- The correlation matrix of the independent variables, say R, can also be used for detecting multicollinearity. The difficulty is that R shows relationships between individual pairs of variables and cannot detect the relationship between each x_j and all other predictors. However, the ith diagonal elements of R^{-1} is the VIF_i.

- The *condition number* of the exogenous matrix X can inform us of linear dependency among the exogenous variables. I.e.

$$\eta = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \geq 1,$$

where λ_j $(j = 1, \ldots, n)$ are the eigenvalues of X.

- Generally the *condition numbers*

$$\eta_j = \sqrt{\frac{\lambda_{\max}}{\lambda_j}}, \quad j = 1, \ldots, n$$

indicate moderate to strong relations if $\eta_j > 30$.

Aside

The BLUE of the standard linear regression model $y = X\beta + \varepsilon$ is given by:

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

If there is multicollinearity, then the condition number η is high. That is, the $(X^T X)^{-1}$ might be singular, or near singular. That is, it will have no solution or provide meaningless estimators.

Example

Consider the highly multicollinear values of the independent variables x_1 and x_2 given in the following table. The dependent variables $y^{(1)}$, $y^{(2)}$ and $y^{(3)}$ may be consider as different samples. They were obtained by adding a $N(0, 0.01)$ pseudo-random numbers to:

$$x_1 + 2x_2$$

and its easily seen that corresponding values of the dependent variables are much alike.

x_1	x_2	$y^{(1)}$	$y^{(2)}$	$y^{(3)}$
2.705	2.695	8.12	8.09	8.09
2.995	3.005	9.01	9.02	9.00
3.255	3.245	9.74	9.75	9.74
3.595	3.605	10.82	10.80	10.79
3.805	3.795	11.38	11.39	11.40
4.145	4.155	12.44	12.44	12.45
4.405	4.395	13.19	13.20	13.19
4.745	4.755	14.27	14.25	14.25
4.905	4.895	14.68	14.70	14.71
4.845	4.855	14.56	14.55	14.54

The VIF of x_1 and x_2 is given by 5868.7.

The condition number of $(x_1 \; x_2)$ is 802.7.

For the $y^{(1)} = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

```
Coefficients:
   Estimate Std. Error t value Pr(>|t|)
x1   0.5926     0.4160   1.425 0.192068
x2   2.4070     0.4159   5.787 0.000411 ***
---
Signif. codes:
Residual standard error: 0.013 on 8 DF
Multiple R-Squared: 1,  Adjusted R-squared:1
F-statistic: 4.19e+06 on 2 and 8 DF, p-value:< 2.2e-16
```

For the $y^{(2)} = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

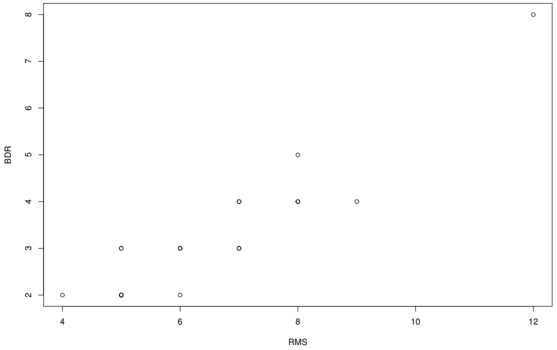
```
Coefficients:
   Estimate Std. Error t value Pr(>|t|)
x1   1.20     0.28     4.27   0.0027 **
x2   1.80     0.28     6.39   0.0002 ***
---
Residual standard error: 0.0089 on 8 DF
Multiple R-Squared: 1, Adjusted R-squared: 1
F-statistic: 9.14e+06 on 2 and 8 DF, p-value: < 2.2e-16
```

For the $y^{(3)} = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

```
Coefficients:
   Estimate Std. Error t value Pr(>|t|)
x1   1.46     0.26   5.71 0.0004 ***
x2   1.54     0.26   6.05 0.0003 ***
---
Residual standard error: 0.008 on 8 DF
Multiple R-Squared: 1, Adjusted R-squared: 1
F-statistic: 1.1e+07 on 2 and 8 DF,  p-value: < 2.2e-16
```

Example (*House prices* model)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.30   5.74   4.06 0.0006 ***
FLR          0.02   0.01   4.15 0.0005 ***
RMS          5.01   1.96   2.55 0.0190 *
BDR         -7.39   2.33  -3.17 0.0049 **
GAR          3.25   1.63   2.00 0.0592 .
ST           9.95   2.77   3.59 0.0018 **
---
Residual standard error: 6.02 on 20 DF
Multiple R-Squared: 0.82, Adjusted R-squared: 0.77
F-statistic: 17.82 on 5 and 20 DF, p-value: 9.2e-07
```



The VIF are given by.

```
vif(House)
 FLR  RMS  BDR  GAR   ST
2.43 7.70 6.40 1.23 1.08
```

There is a fair amount of multicollinearity, particularly involving RMS and BDR.

Calculating the VIF in the *House prices* model

- Fit the regression model:

$$\mathrm{FLR} = \beta_0 + \beta_1\mathrm{RMS} + \beta_2\mathrm{BDR} + \beta_3\mathrm{GAR} + \beta_4\mathrm{ST} + \varepsilon.$$

This gives an $R^2_{\mathrm{FLR}} = 0.589$ and consequently:

$$\mathrm{VIF}_{\mathrm{FLR}} = 1/(1 - R^2_{\mathrm{FLR}}) = 2.43.$$

- Fit the regression model:

$$\mathrm{RMS} = \beta_0 + \beta_1\mathrm{FLR} + \beta_2\mathrm{BDR} + \beta_3\mathrm{GAR} + \beta_4\mathrm{ST} + \varepsilon.$$

to give $R^2_{\mathrm{RMS}} = 0.87$ and $\mathrm{VIF}_{\mathrm{RMS}} = 1/(1 - 0.87) = 7.69$.

- Fit the regression model:

$$\mathrm{BDR} = \beta_0 + \beta_1\mathrm{FLR} + \beta_2\mathrm{RMS} + \beta_3\mathrm{GAR} + \beta_4\mathrm{ST} + \varepsilon.$$

to give $R^2_{\mathrm{BDR}} = 0.84$ and $\mathrm{VIF}_{\mathrm{BDR}} = 1/(1 - 0.84) = 6.25$.

- Fit the regression model:

$$\mathrm{GAR} = \beta_0 + \beta_1\mathrm{FLR} + \beta_2\mathrm{RMS} + \beta_3\mathrm{BDR} + \beta_4\mathrm{ST} + \varepsilon.$$

to give $R^2_{\mathrm{GAR}} = 0.19$ and $\mathrm{VIF}_{\mathrm{GAR}} = 1/(1 - 0.19) = 1.23$.

- Fit the regression model:

$$\mathrm{ST} = \beta_0 + \beta_1\mathrm{FLR} + \beta_2\mathrm{RMS} + \beta_3\mathrm{BDR} + \beta_4\mathrm{GAR} + \varepsilon.$$

to give $R^2_{\mathrm{ST}} = 0.08$ and $\mathrm{VIF}_{\mathrm{ST}} = 1/(1 - 0.08) = 1.08$.

Computing the VIF from the correlation matrix

- The correlation matrix, say R, of the independent variables is given by:

	FLR	RMS	BDR	GAR	ST
FLR	1.00	0.74	0.68	0.40	0.13
RMS	0.74	1.00	0.92	0.30	0.23
BDR	0.68	0.92	1.00	0.24	0.23
GAR	0.40	0.30	0.24	1.00	0.17
ST	0.13	0.23	0.23	0.17	1.00

- The inverse of the correlation matrix, i.e R^{-1} is given by:

	FLR	RMS	BDR	GAR	ST
FLR	2.43	-1.62	-0.07	-0.51	0.17
RMS	-1.62	7.70	-5.87	-0.21	-0.21
BDR	-0.07	-5.87	6.40	0.28	-0.13
GAR	-0.51	-0.21	0.28	1.23	-0.16
ST	0.17	-0.21	-0.13	-0.16	1.08

- Observe that the diagonal elements of R^{-1} are the VIF of the independent variables. That is,
$\mathrm{Diag}(R^{-1}) = (2.43, \quad 7.70, \quad 6.40, \quad 1.23, \quad 1.08) \equiv \mathrm{VIF}.$

- If there is no multicollinearity present, then R, and consequently R^{-1}, have 1 in the diagonal and zero elsewhere. The VIF's show to what extend the variance of an individual variable has been inflated by the presence of multicollinearity.

Summary and example (regression diagnostics)

- It is assumed that there is a linear relationship between years of education (EDU), age (AGE) and salary (SAL). Consider the regression model:

$$\text{SAL}_i = \beta_0 + \beta_1 \text{EDU}_i + \beta_2 \text{AGE}_i + \varepsilon_i.$$

- The data used in the model is given by:

SAL $K	EDU years	AGE years
26.2	12	34
46.5	9	40
28.6	15	37
28.8	16	36
30.4	18	38
34.2	22	44
34.9	24	43

- The coefficient vector β and the data vector y and matrix X in the regression model $y = X\beta + \varepsilon$ are given by:

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, y = \begin{array}{c} \text{SAL} \\ \begin{pmatrix} 26.2 \\ 46.5 \\ 28.6 \\ 28.8 \\ 30.4 \\ 34.2 \\ 34.9 \end{pmatrix} \end{array}, \quad X = \begin{array}{c} (1 \quad \text{EDU} \quad \text{AGE}) \\ \begin{pmatrix} 1 & 12 & 34 \\ 1 & 9 & 40 \\ 1 & 15 & 37 \\ 1 & 16 & 36 \\ 1 & 18 & 38 \\ 1 & 22 & 44 \\ 1 & 24 & 43 \end{pmatrix} \end{array}.$$

- The HAT matrix is given by: $H = X(X^TX)^{-1}X^T$.

$$H = \begin{pmatrix} 0.43 & 0.08 & 0.25 & 0.31 & 0.19 & -0.17 & -0.11 \\ 0.08 & 0.93 & 0.10 & -0.08 & -0.07 & 0.16 & -0.12 \\ 0.25 & 0.10 & 0.19 & 0.21 & 0.17 & 0.03 & 0.05 \\ 0.31 & -0.08 & 0.21 & 0.29 & 0.22 & -0.03 & 0.07 \\ 0.19 & -0.07 & 0.17 & 0.22 & 0.20 & 0.10 & 0.18 \\ -0.17 & 0.16 & 0.03 & -0.03 & 0.10 & 0.47 & 0.43 \\ -0.11 & -0.12 & 0.05 & 0.07 & 0.18 & 0.43 & 0.49 \end{pmatrix}$$

- The diagonal elements of the H matrix, i.e. h_{ii} $(i = 1, \ldots, m)$, denote the leverages of the model: leverages $= \big(0.43,\ 0.93,\ 0.19,\ 0.29,\ 0.20,\ 0.47,\ 0.49\big)$.

- The variance of the predicted values $\hat{y}_i$ is given by $\sigma^2 h_{ii}$.

- The sum of all leverages (i.e. sum of the diagonal elements of $H \equiv \sum_{i=1}^{m} h_{ii}$) is given by the number of variables in the model (including the intercept).

$$\sum_{i=1}^{m} h_{ii} = 0.43 + 0.93 + 0.19 + 0.29 + 0.20 + 0.47 + 0.49 = 3.$$

- This implies, that the total variance of the predicted values of $\hat{y}_i$ is equal to the number of variables in the model times σ^2:

$$\sum_{i=1}^{m} \text{Var}(\hat{y}_i) = \sigma^2 \sum_{i=1}^{m} h_{ii} = n\sigma^2.$$

- If all variances of the predicted values are the same then:

$$\text{Var}(\hat{y}_i) = \sigma^2 h_{ii} = \sigma^2 \frac{n}{m}, \quad \text{or} \quad h_{ii} = \frac{n}{m}.$$

In the example $m = 7$ and $n = 3$. Thus, $n/m = 0.429$.

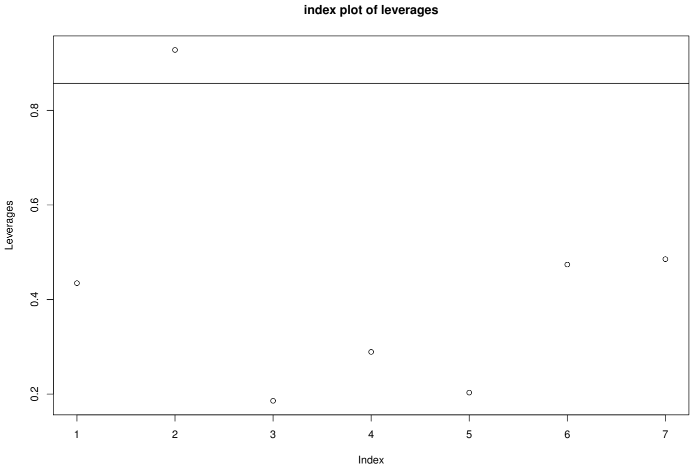
- An observation is influential if its predicted value has *much* bigger variance than the average. Here, twice denotes big. That is,

$$\textit{The ith observation is influential if} \quad h_{ii} > \frac{2n}{m}.$$

- In the (salary) example an observation with a leverage bigger than $2 \times 0.429 = 0.86$ is influential.

- A plot can be used to identify influential observations.

```
            Estimate   SE   t value Pr(>|t|)
(Intercept) -33.27   12.83   -2.59   0.061 .
edu          -1.28    0.27   -4.70   0.009 **
age           2.25    0.39    5.72   0.005 **
---
Residual standard error: 2.69 on 4 DF
Multiple R-Squared: 0.90, Adjusted R-squared: 0.84
F-statistic: 17.2 on 2 and 4 DF, p-value: 0.011
```



- The second observation is influential. Deleting this observation the estimated model change to:

```
           Estimate  SE         t value     Pr(>|t|)
(Intercept) 10     3.439e-14  2.908e+14   <2e-16 ***
edu          0.5   1.272e-15  3.929e+14   <2e-16 ***
age          0.3   1.435e-15  2.091e+14   <2e-16 ***
---
Residual standard error: 3.67e-15 on 3 DF
```

- The residuals are given by:

$$\begin{aligned} e_i &= y_i - \hat{y}_i \\ &= y_i - \hat{\beta}_0 + \hat{\beta}_1 \mathrm{EDU}_i + \hat{\beta}_2 \mathrm{AGE}_i, \quad \text{for } i = 1, \ldots, m. \end{aligned}$$

- In the *Salary* example the residuals are:

$$e = (-1.54 \quad 1.44 \quad -2.04 \quad 1.69 \quad 1.35 \quad -3.20 \quad 2.30)^T$$

- The variance of e_i is given by

$$\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii}), \quad \text{for } i = 1, \ldots, m.$$

- The variance is positive. Thus, $(1 - h_{ii}) > 0$ which implies that

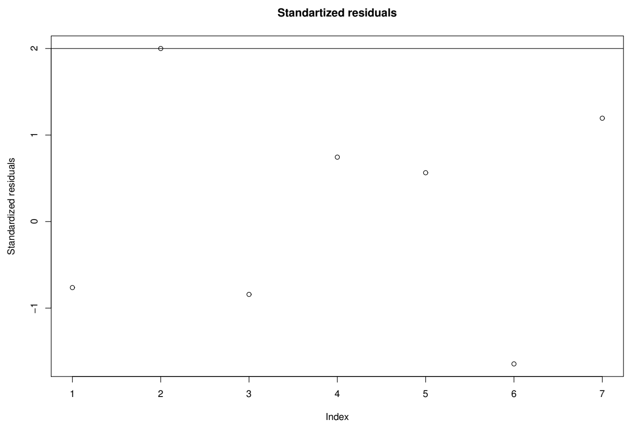
$$0 \leq h_{ii} \leq 1.$$

- The Standardized residuals are given by:

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{(1 - h_{ii})}}.$$

- An observation is an outlier if

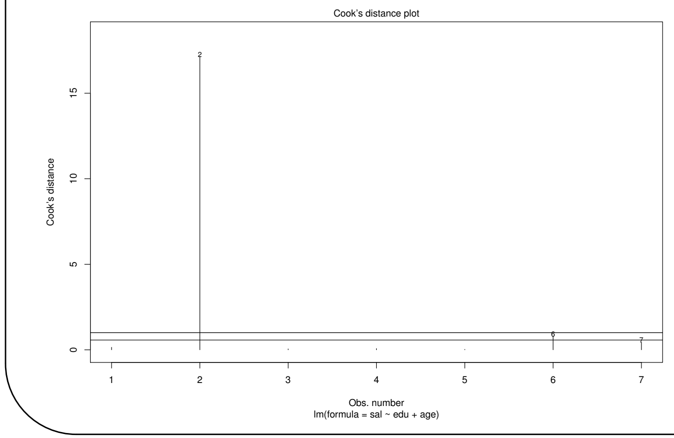
$$\|r_i\| > 2.$$



- Cook's distance, denoted by D_i, is another measure of identifying influential points:

- A cut-off for detecting influential observations are:

$$D_i = 1, \quad \text{or} \quad D_i > 4/m, \quad \text{or} \quad D_i > 4/m.$$



Special Chapters on Artificial Intelligence

Lecture 3. Matrix Algebra and Statistics

Cristian Gatu

[1]Faculty of Computer Science
“Alexandru Ioan Cuza” University of Iași, Romania

MCO, MDS, 2018–2019

In modelling, a lot of problems are linear, or approximated by linear models. Such problems are solved by MATRIX METHODS.

Contents

1. Variance, Sum of squares and Cross Products
2. Definitions and properties of matrices
3. Matrix factorizations
4. Eigenvalues and Eigenvectors
5. Kronecker products
6. Random vector and matrices

Content

Variance, Sum of squares and Cross Products

Definitions and properties of matrices

Matrix factorizations

Eigenvalues and Eigenvectors

Kronecker products

Random vectors and matrices

Variance

- The objective is to account for, or explain, the variation in the data.

- Variance is the most commonly used measure of dispersion in the data.

- Variance directly proportional to the amount of variation or information in the data.

The data below gives two financial ratios, X_1 and X_2, for 12 hypothetical companies.

	Original		Mean-Corrected		Standardize	
	Data		Data		Data	
Firm	X_1	X_2	x_1	x_2	x_1	x_2
1	13	4	7.92	3.83	1.62	1.11
2	10	6	4.92	5.83	1.01	1.69
3	10	2	4.92	1.83	1.01	0.53
4	8	-2	2.92	-2.17	0.60	-0.63
5	7	4	1.92	3.83	0.39	1.11
6	6	-3	0.92	-3.17	0.19	-0.92
7	5	0	-0.08	-0.17	-0.02	-0.05
8	4	2	-1.08	1.83	-0.22	0.53
9	2	-1	-3.08	-1.17	-0.63	-0.34
10	0	-5	-5.08	-5.17	-1.04	-1.49
11	-1	-1	-6.08	-1.17	-1.24	-0.34
12	-3	-4	-8.08	-4.17	-1.65	-1.20
Mean	5.08	0.17	0	0	0	0
SS			262.92	131.67	11	11
Var	23.90	11.97	23.90	11.97	1	1

Mean. Variance

- The MEAN of the jth variable:

$$\boxed{\mu_j = \frac{\sum_{i=1}^{n} X_{ij}}{n}}$$

  where X_{ij} is the ith observation of the jth variable and n is the number of observations.

- The MEAN-CORRECTED jth variable is $x_{ij} = X_{ij} - \mu_j$.
- The VARIANCE of the jth variable:

$$\boxed{s_{jj} = \frac{\sum_{i=1}^{n} x_{ij}^2}{n-1} = \frac{\mathrm{SS}}{\mathrm{df}}}$$

  where SS is the *sum of squares* deviations from the mean and df is the degree of freedom.

Covariance

- Covariation describes the linear relationship, or association, between two variables

- Covariance is a measure of the covariation between two variables X_i and X_j:

$$\boxed{s_{ij} = \frac{\sum_{k=1}^{n} x_{ki} x_{kj}}{n-1} = \frac{\text{SCP}}{\text{df}}}$$

where SCP is the *Sum of the Cross Products* (SCP).

Sum of Squares and Cross Products

- The SS and SCP are summarized in a SUM OF SQUARES AND CROSS PRODUCTS (**SSCP**) matrix.
- The variance and covariances are usually summarized in a *covariance* **S** matrix.
- The **SSCP** and **S** of the two financial ratios are given by:

$$\mathbf{SSCP} = \begin{pmatrix} 262.92 & 136.38 \\ 136.38 & 131.67 \end{pmatrix} \quad \text{and} \quad \mathbf{S} = \begin{pmatrix} 23.90 & 12.40 \\ 12.40 & 11.97 \end{pmatrix}.$$

Note that the matrices are symmetric.

Variance. Covariance

- The variance of a given variable is a measure of its variation in the data. The variances of variables can only be compared if the variables are measured using the same units.

- The Covariance between two variables is a measure of covariation between them. The absolute value of the lower bound covariance is zero implying that the two variables are not linearly associated. However it has no upper bound and this makes it difficult to compare the association between two variables across data sets.

Standardization

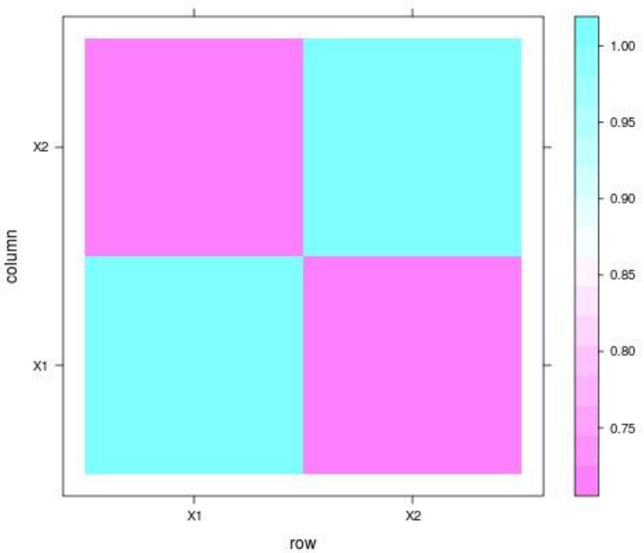
- Standardized data are obtained by dividing the mean-corrected data by the respective standard deviation (square root of variance).
- The variance of the standardized variables is always 1.
- The covariation of standardize variables are always lie between -1 and 1. The value will be:
  - 0 (zero) : no linear relationship between the two variables;
  - -1 (minus one) : a perfect inverse linear relationship;
  - $+1$ (plus one) : a perfect direct linear relationship.

Correlation matrix

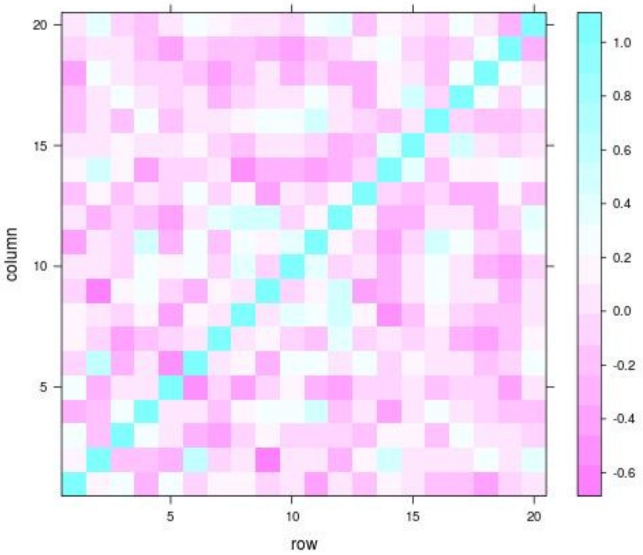
- The covariance of two standardized variables is called the CORRELATION COEFFICIENT.
- The CORRELATION MATRIX ($\mathbf{R}$) is the covariance matrix for standardized data.
- In the example the correlation matrix is:

$$\mathbf{R} = \begin{pmatrix} 1.00 & 0.733 \\ 0.733 & 1.00 \end{pmatrix}.$$

Correlation matrix for the two ratio example



Correlation matrix of 20 variables



Content

Variance, Sum of squares and Cross Products

Definitions and properties of matrices

Matrix factorizations

Eigenvalues and Eigenvectors

Kronecker products

Random vectors and matrices

Matrices

- An $m \times n$ MATRIX A containing $m \times n$ elements has form:

$$A = \begin{array}{@{}c@{}l} \left(\begin{array}{cccccc} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{in} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mj} & \cdots & a_{mn} \end{array}\right) & \begin{array}{l} \\ \\ \\ \leftarrow i\text{th row} \\ \\ \\ \end{array} \end{array}$$

- The subscripts of an element a_{ij} indicates that the element is located at the interception of row i and column j, where $1 \leq i \leq m$ and $1 \leq j \leq n$.

Matrices

- A matrix with one row or one column are called ROW VECTORS or COLUMN VECTORS, respectively.
- A row vector R having n real elements is denoted by $R \in \Re^{1\times n}$ and has the general form $R = (r_1 \ \ldots \ r_n)$.
- A column vector C having m real elements is denoted by $C \in \Re^{m\times 1}$ and has the general form

$$C = \begin{pmatrix} c_1 \\ \vdots \\ c_m \end{pmatrix}.$$

- Generally a C m-elements real vector will be assumed to be a column vector and denoted by $C \in \Re^{m}$.

Special types of matrices

- Square matrix: an $m \times n$ matrix is square if $m = n$.

- Identity (or unit) matrix: I_m.

- Transposed of a matrix: if $A = [a_{ij}] \in \Re^{m \times n}$, $B = [b_{ij}] \in \Re^{n \times m}$ and $b_{ji} = a_{ij}$ then $B = A^T$.

- Symmetric matrix: $A = A^T$.

- Upper triangular matrix: $U = [a_{ij}] \in \Re^{m \times n}$ s.t. $\forall i > j,\ u_{ij} = 0$

- Lower triangular matrix: $L = [a_{ij}] \in \Re^{m \times n}$ s.t. $\forall i - j < m - n,\ l_{ij} = 0$.

Matrix operations

- Two matrices can be ADDED or SUBTRACTED (element by element) iff they have the same dimension.
- The MULTIPLICATION OF A SCALAR BY A MATRIX is equivalent into multiplying each element of the matrix by the scalar.
- The INNER PRODUCT is an operation between a row and a column vector (in this order). It is computed by multiplying corresponding elements in the two vectors and algebraically summing.
- MATRIX MULTIPLICATION. Given $A \in \Re^{m_a \times n_a}$ and $B \in \Re^{m_b \times n_b}$ the matrix product $C = A\,B$ is defined iff $n_a = m_b$. The element c_{ij} is defined to be the inner product of row i in matrix A and column j in matrix B.

Partitioned matrices

- A partitioned matrix contains sub-matrices as elements.
- E.g. consider the partitioning of $A, B \in \Re^{m \times n}$ as:

$$A = \begin{array}{cc} \begin{array}{ccc} n_1 & \quad & n_N \end{array} & \\ \begin{pmatrix} A_{11} & \dots & A_{1N} \\ \vdots & & \vdots \\ A_{M1} & \dots & A_{MN} \end{pmatrix} & \begin{array}{c} m_1 \\ \\ m_M \end{array} \end{array} \text{ and } B = \begin{array}{cc} \begin{array}{ccc} n_1 & \quad & n_N \end{array} & \\ \begin{pmatrix} B_{11} & \dots & B_{1N} \\ \vdots & & \vdots \\ B_{M1} & \dots & B_{MN} \end{pmatrix} & \begin{array}{c} m_1 \\ \\ m_M \end{array} \end{array},$$

where $n = \sum_{i=1}^{N} n_i$ and $m = \sum_{i=1}^{M} m_i$.

- addition and multiplication of partitioned matrices.

$$A + B = \begin{pmatrix} A_{11} + B_{11} & \dots & A_{1N} + B_{MN} \\ \vdots & & \vdots \\ A_{M1} + B_{M1} & \dots & A_{MN} + B_{MN} \end{pmatrix}.$$

Rank of a matrix

- The number of linearly independent columns of a matrix is called COLUMN RANK, hereafter RANK. It will be denoted by $\text{rank}(A)$.

- The square matrix $A \in \Re^{n \times n}$ is said to be *non-singular* if the $\text{rank}(A) = n$. Otherwise it is called *singular*.

- Properties
  1. $\text{rank}(A) = \text{rank}(A^T)$.
  2. $\text{rank}(A) = \text{rank}(A^T A) = \text{rank}(AA^T)$.
  3. The rank of A is unchanged by pre- or postmultiplication of A by a non-singular matrix.

Trace of a matrix

- For a square matrix $A = [a_{ij}] \in \Re^{n\times n}$ the sum of its diagonal elements is called its trace, i.e.

$$\boxed{\text{trace}(A) = \sum_{i=1}^{n} a_{ii}}.$$

- Properties

1. $\text{trace}(A) = \text{trace}(A^T)$.
2. $\text{trace}(AB) = \text{trace}(BA)$.
3. $\text{trace}(ABC) = \text{trace}(BCA) = \text{trace}(CAB)$.
4. $\text{trace}(A + B) = \text{trace}(B + A) = \text{trace}(A) + \text{trace}(B)$.
5. $\text{trace}\left(\sum_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} \text{trace}(A_i)$.
6. $\text{trace}(\kappa\, A) = \kappa\ \text{trace}(A)$.

Matrix properties

- For any two matrices A and B, it CANNOT be stated that $AB = BA$.
- If A is an $m \times n$ matrix, then $I_m A = A I_n = A$.
- $(AB)^T = B^T A^T$. Generally: $(A_1 A_2 \cdots A_n)^T = A_n^T \cdots A_2^T A_1^T$.

Inverse of a matrix

- The relationship between a square matrix A and its inverse, denoted by A^{-1} (inverse of A), is that:

$$\boxed{A^{-1}A = AA^{-1} = I}.$$

- Note that
  - — The matrix A must be square.
  - — The dimensions of A and A^{-1} are the same.
  - — Only non-singular matrices have an inverse.
- For $|A| \neq 0$, the inverse of A is given by:

$$\boxed{A^{-1} = \frac{1}{|A|} A_C^T}.$$

Gaussian reduction procedure

- Consider the $m \times m$ matrix A. Construct the augmented matrix $(A \mid I_m)$.

- The Gaussian elimination method transforms $(A \mid I_m)$ to $(I_m \mid A^{-1})$ by applying two basic operations:
  1. Rows can be multiplied by a non zero constant; and
  2. non zero multiples of one row can be added to another row.

Properties of the inverse

- The inverse of a symmetric matrix is also symmetric.
- $(A^T)^{-1} = (A^{-1})^T = A^{-T}$.
- Let $A_1, \ldots, A_n \in \Re^{n \times n}$. Then, $(A_1 A_2 \cdots A_n)^{-1} = (A_n^{-1} \cdots A_2^{-1} A_1^{-1})$.
- If c is a non zero scalar, then $(cA)^{-1} = \frac{1}{c} A^{-1}$.
- The inverse of a diagonal matrix is a diagonal matrix consisting of the reciprocals of the original elements.
- The inverse of a triangular matrix is also triangular.

System of equations

Consider the $n \times n$ system of equations having the form:

$$\begin{array}{lllll}
a_{11}x_1 & +a_{12}x_2 & +\ldots & +a_{1n}x_n & = b_1 \\
a_{21}x_1 & +a_{22}x_2 & +\ldots & +a_{2n}x_n & = b_2 \\
\vdots & \vdots & & \vdots & \vdots \\
a_{n1}x_1 & +a_{n2}x_2 & +\ldots & +a_{nn}x_n & = b_n
\end{array}$$

can be written in a matrix form as

$$\begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \tag{1}$$

$$\text{or} \quad Ax = b \tag{2}$$

System of equations

- Assume that the equations are linear independent, that is, A is not singular (it has inverse).
- Premultiply both sides of (2) by A^{-1} it gives:

$$A^{-1}Ax = A^{-1}b \quad \text{or} \quad x = A^{-1}b$$

  since $A^{-1}A\,x = I_n\,x = x$.
- Thus, the solution of (1) is given by $\boxed{x = A^{-1}b}$.

Orthogonal matrices

- A square matrix $Q \in \Re^{m \times m}$ is orthogonal iff

$$\boxed{Q^T Q = QQ^T = I_m}.$$

- Notice that the inverse of Q is given by Q^T.

- Examples of orthogonal matrices:

$$I_m, \quad \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix}.$$

Orthogonal matrices. Property

It preserves the norm (inner product) of a vector.
That is, If $z = Qx$ and Q is orthogonal, then $z^T z = x^T x$.

Note $z^T z = (Qx)^T (Qx) = x^T Q^T Qx = x^T Ix = x^T x$.

Example

$$x = \begin{pmatrix} -1 \\ 3 \end{pmatrix} \quad \text{and} \quad Q = \begin{pmatrix} 0.5 & 0.866 \\ -0.866 & 0.5 \end{pmatrix}.$$

$$z = Qx = \begin{pmatrix} 2.098 \\ 2.366 \end{pmatrix} \quad \text{and} \quad x^T x = 10 = z^T z.$$

Content

Variance, Sum of squares and Cross Products

Definitions and properties of matrices

Matrix factorizations

Eigenvalues and Eigenvectors

Kronecker products

Random vectors and matrices

Cholesky Decomposition

The CHOLESKY DECOMPOSITION of a symmetric positive definite $n \times n$ matrix A, is given by

$$\boxed{A = LL^T},$$

where $L \in \Re^{n \times n}$ is lower triangular and non-singular. E.g.

$$\begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{12} & A_{22} & A_{23} \\ A_{13} & A_{23} & A_{33} \end{pmatrix} = \begin{pmatrix} L_{11} & 0 & 0 \\ L_{21} & L_{22} & 0 \\ L_{31} & L_{32} & L_{33} \end{pmatrix} \begin{pmatrix} L_{11} & L_{21} & L_{31} \\ 0 & L_{22} & L_{32} \\ 0 & 0 & L_{33} \end{pmatrix}.$$

Cholesky Decomposition. Example

Let $\quad A = \begin{pmatrix} 5 & 2 & 3 \\ 2 & 10 & 1 \\ 3 & 1 & 2 \end{pmatrix}$

The Cholesky Decomposition of $A = LL^T$ is given by:

$$\begin{pmatrix} 2.24 & 0 & 0 \\ 0.89 & 3.03 & 0 \\ 1.34 & -0.07 & 0.44 \end{pmatrix} \begin{pmatrix} 2.24 & 0.89 & 1.34 \\ 0 & 3.03 & -0.07 \\ 0 & 0 & 0.44 \end{pmatrix}$$

Cholesky Decomposition. Application

Solve the matrix problem $Ax = b$, where A is symmetric and has Cholesky decomposition $A = LL^T$.

Notice that $L(L^T x) = b$ is equivalent to $Lz = b$, where $L^T x = z$. That is, the solution of $Ax = b$ comes in three steps:

1. Compute the Cholesky decomposition $A = LL^T$.
2. Solve the lower-triangular system $Lz = b$ for z.
3. Solve the upper-triangular system $L^T x = z$ for x.

Cholesky Decomposition. Application

Example

Solve $Ax = b$, where $A = \begin{pmatrix} 5 & 2 & 3 \\ 2 & 10 & 1 \\ 3 & 1 & 2 \end{pmatrix}$ and $b = \begin{pmatrix} 7 \\ -16 \\ 5 \end{pmatrix}$.

1. $A = LL^T$, where $L = \begin{pmatrix} 2.24 & 0 & 0 \\ 0.89 & 3.03 & 0 \\ 1.34 & -0.07 & 0.44 \end{pmatrix}$
2. Solve $Lz = b$ which gives $z = \begin{pmatrix} 3.13 \\ 6.19 \\ 0.88 \end{pmatrix}$
3. Solve $L^T x = z$ which gives $x = \begin{pmatrix} 1 \\ -2 \\ 2 \end{pmatrix}$.

QR decomposition

Let $A \in \Re^{m \times n}$ $(m \geq n)$ have full column rank.

The QR DECOMPOSITION of A has the form:

$$\boxed{A = QR},$$

where $R \in \Re^{n \times n}$ is upper triangular and $Q \in \Re^{m \times m}$ is orthogonal.

QR decomposition. Example

Let $A = Q\begin{pmatrix} R \\ 0 \end{pmatrix}$, where $A \in \Re^{5\times 3}$, $Q \in \Re^{5\times 5}$ is orthogonal and $R \in \Re^{3\times 3}$ is upper-triangular.

$$A = \begin{pmatrix} -8 & -2 & 8 \\ -9 & 7 & 3 \\ -13 & -14 & 17 \\ 4 & 3 & -13 \\ -4 & 1 & 16 \end{pmatrix}, \quad R = \begin{pmatrix} 18.6 & 7.69 & -23.00 \\ 0 & -14.14 & 5.60 \\ 0 & 0 & 15.04 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \text{ and}$$

$$Q = \begin{pmatrix} -0.43 & -0.09 & -0.09 & 0.55 & -0.70 \\ -0.48 & -0.76 & -0.26 & -0.28 & 0.20 \\ -0.70 & 0.61 & -0.17 & -0.05 & 0.33 \\ 0.22 & -0.10 & -0.50 & 0.68 & 0.48 \\ -0.22 & -0.19 & 0.80 & 0.38 & 0.35 \end{pmatrix}.$$

$$Q^TQ = QQ^T = I_5 \quad \text{and} \quad A = QR.$$

QR decomposition. Application

Solve the matrix problem $Ax = b$ using the QR decomposition, where $A \in \Re^{n \times n}$ is non singular.

Let $A = QR$.
The system $Ax = b$ can be written as $QRx = b$.
Premultiply both sides of the system by Q^T it gives:

$$Q^T QRx = Q^T b$$

Since $Q^T Q = I_n$ the latter is equivalent to

$$Rx = Q^T b.$$

QR decomposition. Example

$$A = \begin{pmatrix} 5 & 2 & 3 \\ 2 & 10 & 1 \\ 3 & 1 & 2 \end{pmatrix} \text{ and } b = \begin{pmatrix} 7 \\ -16 \\ 5 \end{pmatrix}.$$

$$A = QR = \begin{pmatrix} -0.81 & 0.27 & -0.52 \\ -0.32 & -0.95 & 0.02 \\ -0.49 & 0.18 & 0.85 \end{pmatrix} \begin{pmatrix} -6.16 & -5.35 & -3.73 \\ 0 & -8.74 & 0.23 \\ 0 & 0 & 0.17 \end{pmatrix}$$

$$Q^T b = \begin{pmatrix} -2.92 \\ 17.93 \\ 0.33 \end{pmatrix} \quad \text{and} \quad Rx = Q^T b \quad \text{gives} \quad x = \begin{pmatrix} 1 \\ -2 \\ 2 \end{pmatrix}.$$

Computing the QRD

- Givens Rotations
- Householder transformations
- Gram-Schmidt process

Singular Value Decomposition (SVD)

Let $A \in \Re^{m \times n}$ be a matrix of rank k.
The SINGULAR VALUE DECOMPOSITION (SVD) of A is given by:

$$\boxed{A = Q \Sigma P^T},$$

- where $Q \in \Re^{m \times m}$ and $P \in \Re^{n \times n}$ are orthogonal,
- $\Sigma = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}$

SVD

- $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k > 0 \quad$ and $\quad \sigma_{k+1} = \ldots \sigma_n = 0.$

- The rank of A is k.

- The σ_i is called the ith singular value of A.

- If $Q = (q_1, \ldots q_m)$ and $P = (p_1, \ldots p_n)$, then q_i and p_i are called the left and right singular vectors associated with σ_i $(i = 1, \ldots, k)$.

- The ratio $\kappa(A) = \sigma_1 / \sigma_n$ is called the condition number of A.

SVD. Example

$$A = \begin{pmatrix} -6 & -12 & 8 \\ 2 & 12 & -11 \\ -6 & -17 & 10 \\ 19 & 3 & 6 \\ -9 & 6 & 15 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 31.71 & 0 & 0 \\ 0 & 19.80 & 0 \\ 0 & 0 & 16.99 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$Q = \begin{pmatrix} -0.49 & -0.05 & -0.11 & 0.51 & -0.70 \\ 0.49 & 0.29 & 0.02 & 0.80 & 0.22 \\ -0.63 & -0.15 & -0.23 & 0.26 & 0.68 \\ 0.23 & -0.93 & 0.20 & 0.19 & 0.01 \\ -0.27 & 0.15 & 0.94 & 0.06 & 0.09 \end{pmatrix}, P = \begin{pmatrix} 0.46 & -0.88 & -0.15 \\ 0.68 & 0.23 & 0.70 \\ -0.58 & -0.42 & 0.70 \end{pmatrix}.$$

The Condition number of A is given by $\sigma_1/\sigma_3 = 31.77/16.99 = 1.87$.
Consider the matrices:

$$A_0 = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 4 & 1 \\ 3 & 6 & 0 \\ 4 & 8 & 2 \end{pmatrix}, \; A_1 = \begin{pmatrix} 1 & 2.01 & 0 \\ 2 & 3.99 & 1 \\ 3 & 6 & 0 \\ 4 & 8 & 2 \end{pmatrix}, \; A_2 = \begin{pmatrix} 1 & 2.1 & 0 \\ 2 & 3.9 & 1 \\ 3 & 6 & 0 \\ 4 & 8 & 2 \end{pmatrix}, \; A_3 = \begin{pmatrix} 1 & 1 & 0 \\ 2 & 4 & 1 \\ 3 & 9 & 0 \\ 4 & 16 & 2 \end{pmatrix}.$$

Cond(A_0)=8.82e+16, Cond(A_1)=2124.5, Cond(A_2)=213.02,
Cond(A_3)=17.77, and Cond(I_n)=1.

Content

Variance, Sum of squares and Cross Products

Definitions and properties of matrices

Matrix factorizations

Eigenvalues and Eigenvectors

Kronecker products

Random vectors and matrices

The Eigenvalue problem

- Let A be a square matrix of order $n \times n$, $x \neq 0$ is an n-element column vector and λ is a scalar.
- The EIGENVALUE PROBLEM: Solve

$$\boxed{Ax = \lambda x}.$$

- The solution come in pairs: to each λ corresponds an x vector.
- The λ's are known as eigenvalues (or latent, or characteristic roots).
- The x's as eigenvectors (or latent, or, characteristic vectors).

The Eigenvalue problem

- In matrix format the Eigenvalue problem can be written as:
$$\boxed{(A - \lambda I_n)x = 0}$$

- In order for $x \neq 0$ it implies that
$$\boxed{|A - \lambda I_n| = 0}.$$

- The latter is known as the *characteristic equation* for A. It gives a polynomial equation in the unknown λ.

Example

Let $A = \begin{pmatrix} 1 & 0 \\ 1 & 3 \end{pmatrix}$ so that $A - \lambda I_2 = \begin{pmatrix} 1-\lambda & 0 \\ 1 & 3-\lambda \end{pmatrix}$.
Now, $|A - \lambda I_2| = (1-\lambda)(3-\lambda)$.
Thus, $\lambda_1 = 1$ and $\lambda_2 = 3$ are the eigenvalues of A.

For the eigenvalue $\lambda_1 = 1$ we have $Ax = \lambda_1 x$:

$$\begin{pmatrix} 1 & 0 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \text{or} \quad \begin{aligned} x_1 &= x_1 \\ x_1 &= -2x_2. \end{aligned}$$

Thus, an eigenvector of A corresponding to the eigenvalue $\lambda_1 = 1$ is given by $x = (-2 \quad 1)^T$. Normalizing x, i.e. dividing each of its entries by $\sqrt{x^T x}$, it gives the eigenvector

$$\frac{1}{\sqrt{5}} \begin{pmatrix} -2 \\ 1 \end{pmatrix}.$$

An eigenvector associated with the eigenvalue of $\lambda_2 = 3$ is given by $(0 \quad 1)^T$.

Properties of eigenvalues and eigenvectors

Given an $m \times m$ SYMMETRIC matrix, e.g. the variance-covariance matrix: $A = \begin{pmatrix} 5 & 2 & 3 \\ 2 & 10 & 1 \\ 3 & 1 & 2 \end{pmatrix}$

- **The eigenvalues are real.** The eigenvalue of A are given by:
  $\lambda_1 = 0.14, \quad \lambda_2 = 5.70 \quad \text{and} \quad \lambda_3 = 11.16.$

Properties of eigenvalues and eigenvectors

- **Eigenvectors corresponding to dinstinct eigenvalues are pairwise orthogonal**[1]. I.e. if x_1 and x_2 are the eigenvectors corresponding to the eigenvalues λ_1 and λ_2 $(\lambda_1 \neq \lambda_2)$, then $x_1^T x_2 = 0$.

  The eigenvectors of A are given by the columns of $X = (x_1, x_2, x_3)$, where

  $$X = \begin{pmatrix} -0.532 & 0.747 & 0.400 \\ 0.022 & -0.459 & 0.888 \\ 0.847 & 0.481 & 0.228 \end{pmatrix} \quad \text{and} \quad X^T X = X X^T = I_3.$$

[1] Notice that $Ax_1 = \lambda_1 x_1$, and after premultiplicationn by x_2^T it gives $x_2^T A x_1 = \lambda_1 x_2^T x_1$. Similarly, $x_1^T A x_2 = \lambda_2 x_1^T x_2$. Since $x_2^T A x_1 = x_1^T A x_2$ it follows that $\lambda_1 x_2^T x_1 = \lambda_2 x_1^T x_2$ and thus, $x_1^T x_2 = 0$.

Properties of eigenvalues and eigenvectors

- **The orthogonal matrix of eigenvectors diagonalizes**[2] A. That is,

$$\boxed{X^T A X = \Lambda},$$

where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_m)$ and $X = (x_1 \ldots x_m)$.

$$X^T A X = \begin{pmatrix} 0.14 & 0 & 0 \\ 0 & 5.70 & 0 \\ 0 & 0 & 11.16 \end{pmatrix} = \Lambda$$

- **The matrices A and A^T have the same eigenvalues.**
- **The matrix A is singular if one of its eigenvalues is zero.**
- **The rank of A is equal to the number of non-zero eigenvalues.**

[2] The Eigenvalue problem in matrix form is equivalent to $AX = X\Lambda$. Premultiplying by X^T it gives $X^T AX = X^T X\Lambda$ which is equivalent to $X^T AX = \Lambda$ since $X^T X = I_m$.

Properties of eigenvalues and eigenvectors

- $A^2 = AA = X\Lambda^2 X^T$ and generally $A^n = X\Lambda^n X^T$.

$$A^2 = \begin{pmatrix} 38 & 33 & 23 \\ 33 & 105 & 18 \\ 23 & 18 & 14 \end{pmatrix} \quad \text{and} \quad \Lambda^2 = \begin{pmatrix} 0.02 & 0 & 0 \\ 0 & 32.51 & 0 \\ 0 & 0 & 124.47 \end{pmatrix}.$$

- $A^{-1} = X\Lambda^{-1}X^T$ since $(X\Lambda X^T)^{-1} = X\Lambda^{-1}X^T$.

$$A^{-1} = \frac{1}{9}\begin{pmatrix} 19 & -1 & -28 \\ -1 & 1 & 1 \\ -28 & 1 & 46 \end{pmatrix} \text{ and } \Lambda^{-1} = \begin{pmatrix} 7.07 & 0 & 0 \\ 0 & 0.18 & 0 \\ 0 & 0 & 0.09 \end{pmatrix}.$$

Equiv. of the SVD of A and Eigensystem of A^TA

Consider the SVD of $A \in \Re^{m\times n}$: $\quad A = Q\Sigma P^T$, where Q and P have orthogonal columns, and $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_n)$. Now, $\quad A^TA = (P\Sigma Q^T)(Q\Sigma P^T) = P\Sigma^2 P^T$, or

$$\boxed{P^TA^TAP = \Sigma^2}.$$

Thus, the SVD of A provides:

- The eigenvectors P of the symmetric A^TA
- The diagonal elements of Σ are the positive square roots of the eigenvalues of A^TA. I.e. $\lambda_1 = \sigma_1^2, \dots, \lambda_n = \sigma_n^2$.

Equiv. of the SVD of A and Eigensystem of $A^T A$

Let $A = \begin{pmatrix} 5 & 2 & 3 \\ 2 & 10 & 1 \\ 3 & 1 & 2 \end{pmatrix}$ such that $A^T A = \begin{pmatrix} 38 & 33 & 23 \\ 33 & 105 & 18 \\ 23 & 18 & 14 \end{pmatrix}$.

The singular values of A are: $\sigma_1 = 11.16$, $\sigma_2 = 5.70$ and $\sigma_3 = 0.14$.

The eigenvalues of $A^T A$ are: $\lambda_1 = 124.47$, $\lambda_2 = 32.51$ and $\lambda_3 = 0.02$.

Quadratic forms and definite matrices

Consider the quadratic form $q = x^T Ax$, where A is a symmetric matrix and $x \neq 0$. E.g. if $A \in \Re^{2\times 2}$, then

$$\boxed{q = x^T Ax = a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2}.$$

- If $x^T Ax > 0$, then the quadratic form is said to be positive definite. *In this case all the eigenvalues of A are positive.*
  E.g. Let $S = \begin{pmatrix} 5 & 2 \\ 2 & 10 \end{pmatrix}$ such that $\Lambda = \begin{pmatrix} 10.70 & 0 \\ 0 & 4.29 \end{pmatrix}$.

Quadratic forms and definite matrices

- If $x^T Ax \geq 0$, then the quadratic form is said to be positive (or nonnegative) semidefinite. *In this case all the eigenvalues of A are positive or zero.*
  E.g. Let $S = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$ such that $\Lambda = \begin{pmatrix} 5 & 0 \\ 0 & 0 \end{pmatrix}$.

- If $x^T Ax < 0$, then the quadratic form is said to be negative definite. *In this case all the eigenvalues of A are negative.*

- If $x^T Ax \leq 0$, then the quadratic form is said to be negative (or nonpositive) semidefinite. *In this case all the eigenvalues of A are negative or zero.*

Content

Variance, Sum of squares and Cross Products

Definitions and properties of matrices

Matrix factorizations

Eigenvalues and Eigenvectors

Kronecker products

Random vectors and matrices

Kronecker products

- A calculation that helps condense the notation when dealing with sets of regression models are the *Kronecker product* and *vector operator*.

- The KRONECKER PRODUCT of the two matrices $A = [a_{ij}] \in \Re^{m \times n}$ and $B = [b_{ij}] \in \Re^{p \times q}$ is defined by:

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \dots & a_{mn}B \end{pmatrix}$$

- Notice that $A \otimes B$ has dimension $mp \times nq$.

Kronecker products. Example

Let $A = \begin{pmatrix} 3 & 0 \\ 5 & 2 \end{pmatrix}$ and $B = \begin{pmatrix} 1 & 4 \\ -1 & 0 \\ -2 & 1 \end{pmatrix}$:

$$A \otimes B = \begin{pmatrix} 3\begin{pmatrix} 1 & 4 \\ -1 & 0 \\ -2 & 1 \end{pmatrix} & 0\begin{pmatrix} 1 & 4 \\ -1 & 0 \\ -2 & 1 \end{pmatrix} \\ 5\begin{pmatrix} 1 & 4 \\ -1 & 0 \\ -2 & 1 \end{pmatrix} & 4\begin{pmatrix} 1 & 4 \\ -1 & 0 \\ -2 & 1 \end{pmatrix} \end{pmatrix} = \left(\begin{array}{rr|rr} 3 & 12 & 0 & 0 \\ -3 & 0 & -0 & 0 \\ -6 & 3 & -0 & 0 \\ \hline 5 & 20 & 2 & 8 \\ -5 & 0 & -2 & 0 \\ -10 & 5 & -4 & 2 \end{array}\right).$$

$$A \otimes I_2 = \begin{pmatrix} 3\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} & 0\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \\ 5\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} & 2\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \end{pmatrix} = \left(\begin{array}{cc|cc} 3 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ \hline 5 & 0 & 2 & 0 \\ 0 & 5 & 0 & 2 \end{array}\right).$$

$$I_2 \otimes A = \begin{pmatrix} A & 0 \\ 0 & A \end{pmatrix} = \left(\begin{array}{cc|cc} 3 & 0 & 0 & 0 \\ 5 & 2 & 0 & 0 \\ \hline 0 & 0 & 3 & 0 \\ 0 & 0 & 5 & 2 \end{array}\right).$$

Direct sum of matrices

- Given the set of matrices $\{A_1, \ldots, A_G\}$ the DIRECT SUM of matrices is defined by:

$$\bigoplus_{i=1}^{G} A_i = \mathrm{diag}(A_1, \ldots, A_G) = \begin{pmatrix} A_1 & & \\ & \ddots & \\ & & A_G \end{pmatrix}.$$

- Notice that the matrices $A_1, \ldots, A_G$ can have different dimensions.

- In the event where the matrices are of the same (A) then:

$$\bigoplus_{i=1}^{G} A_i = I_G \otimes A.$$

Vector operator

Let the $m \times n$ matrix $Y = (y_1 \dots y_n)$ where $y_i \in \Re^m$ is the ith column of Y. The vec($\cdot$) operator stacks the columns of Y one under the other. That is,

$$\boxed{\mathrm{vec}(Y) = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}.$$

E.g. If $Y = \begin{pmatrix} 1 & 4 \\ 1 & 0 \\ 2 & 1 \end{pmatrix}$, then $\mathrm{vec}(Y) = \begin{pmatrix} 1 & 1 & 2 & 4 & 0 & 1 \end{pmatrix}^T$.

Vector operator

Given the set of vectors $\{y_i\}_G = \{y_1, \ldots, y_G\}$ the $\mathrm{vec}(\cdot)$ operator stacks the vectors one under the other:

$$\boxed{\{\mathrm{vec}(y_i)_G\} = \begin{pmatrix} y_1 \\ \vdots \\ y_G \end{pmatrix}}.$$

Properties

Assuming appropriate dimensions the following properties exist:

- $(A \otimes B)(C \otimes D) = (AC \otimes BD)$.
- $(A \otimes B)^T = (A^T \otimes B^T)$.
- $(A \otimes B)^{-1} = (A^{-1} \otimes B^{-1})$.
- $|A \otimes B| = |A|^M |B|^N$.
- $(A \otimes B)\text{vec}(C) = \text{vec}(BCA^T)$.

Vector and Frobenius norms

The p-norm of $x \in \Re^n$ is defined as:

$$\boxed{\|x\|_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{\frac{1}{p}}, \quad \text{where} \quad p \geq 1.}$$

Important norms:

- $\|x\|_1 = (|x_1| + \cdots + |x_n|)$.
- $\|x\|_2 = (|x_1|^2 + \cdots + |x_n|^2)^{\frac{1}{2}} = \sqrt{x^T x}$.
  The 2-norm is also known as EUCLIDIAN NORM.
  Hereafter $\|\cdot\|$ will denote the Euclidian norm.
- $\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$.

Given a matrix $A = [a_{ij}] \in \Re^{m \times n}$ the FROBENIUS NORM of A is given by:

$$\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2}.$$

Absolute and relative error

Suppose $\hat{x} \in \Re^n$ is an approximation of $x \in \Re^n$. Then:

- Absolute error in $\hat{x}$: $\|\hat{x} - x\|$.
- Relative error in $\hat{x}$: $\|\hat{x} - x\| \Big/ \|x\|$.

Content

Variance, Sum of squares and Cross Products

Definitions and properties of matrices

Matrix factorizations

Eigenvalues and Eigenvectors

Kronecker products

Random vectors and matrices

Random vectors and matrices

A RANDOM VECTOR (RANDOM MATRIX) is a vector (matrix) whose elements are random variables. The expected value of a random matrix consists of the expected values of each element. That is, if $X = [X_{ij}] \in \Re^{m\times n}$, then

$$\boxed{E(X) = \begin{pmatrix} E(X_{11}) & \dots & E(X_{1n}) \\ \vdots & & \vdots \\ E(X_{m1}) & \dots & E(X_{mn}) \end{pmatrix}},$$

where

$$E(X_{ij}) = \begin{cases} \int_{-\infty}^{\infty} x_{ij} f_{ij}(x_{ij}) dx_{ij} & \text{If } X_{ij} \text{ is continues random variable with pdf } f_{ij}(x_{ij}) \\ \sum\limits_{\text{all } x_{ij}} x_{ij} p_{ij}(x_{ij}) & \text{If } X_{ij} \text{ is discrete r.v. with probability function } p_{ij}(x_{ij}) \end{cases}$$

Example

Consider the random vector $X^T = (X_1 \quad X_2)$, where X_1 and X_2 have, respectively, the following probability functions:

x_1	-1	0	1
$p_1(x_1)$	0.3	0.3	0.4

x_2	0	1
$p_2(x_2)$	0.8	0.2

Thus, $E(X) = (E(X_1) \quad E(X_2))^T = (0.1 \quad 0.2)^T$.

Mean vectors and covariance matrices

Consider the random vector $X = [X_i] \in \Re^n$, where X_i has mean $\mu_i = E(X_i)$ and variance $\sigma_i^2 = E(X_i - \mu_i)^2$, where $i = 1, \ldots, n$.

The means of the vector X is given by:

$$\boxed{E(X) = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_n) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} = \mu.}$$

Mean vectors and covariance matrices

Specifically,

$$\mu_i = \begin{cases} \int_{-\infty}^{\infty} x_i f_i(x_i) dx_i & \text{if } X_i \text{ is continues random variable} \\ & \text{with pdf } f_i(x_i) \\ \sum\limits_{\text{all } x_i} x_i p_i(x_i) & \text{if } X_i \text{ is discrete random variable} \\ & \text{with probility function } p_i(x_i) \end{cases}$$

and

$$\sigma_i^2 = \begin{cases} \int_{-\infty}^{\infty} (x_i - \mu_i)^2 f_i(x_i) dx_i & \text{if } X_i \text{ is continues random variable} \\ & \text{with pdf } f_i(x_i) \\ \sum\limits_{\text{all } x_i} (x_i - \mu_i)^2 p_i(x_i) & \text{if } X_i \text{ is discrete random variable} \\ & \text{with probility function } p_i(x_i) \end{cases}$$

Mean vectors and covariance matrices

- The behavior of any pair of random variables, such as X_i and X_k, is described by their joint probability function and a the covariance σ_{ik}.

- The covariance σ_{ik} is measure of a linear associassion between the two variables and is given by:

$$\boxed{\sigma_{ik} = E(X_i - \mu_i)(X_k - \mu_k).}$$

Mean vectors and covariance matrices

If X_i and X_k are continues random variables with joint density function $f_{ik}(x_i, x_k)$, then

$$\boxed{\sigma_{ik} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_i - \mu_i)(x_k - \mu_k)f_{ik}(x_i, x_k)dx_i dx_k.}$$

If X_i and X_k are discrete random variables with joint probability function $p_{ik}(x_i, x_k)$, then

$$\boxed{\sigma_{ik} = \sum_{\text{all } x_i}\sum_{\text{all } x_k}(x_i - \mu_i)(x_k - \mu_k)p_{ik}(x_i, x_k).}$$

Generally, the collective behavior of the n random variables $X_1, \ldots, X_n$, or equivalently the random vector $X^T = (X_1 \ldots X_n)$ is described by the a joint probability density function $f(x_1, \ldots, x_n) = f(x)$ and their covariance matrix.

Mean vectors and covariance matrices

The covariances of the vector X is given by:

$$Var(X) = \Sigma = E(X-\mu)(X-\mu)^T$$

$$= \mathrm{E}\begin{pmatrix} (X_1-\mu_1)^2 & (X_1-\mu_1)(X_2-\mu_2) & \dots & (X_1-\mu_1)(X_n-\mu_n) \\ (X_2-\mu_2)(X_1-\mu_1) & (X_2-\mu_2)^2 & \dots & (X_2-\mu_2)(X_n-\mu_n) \\ \vdots & \vdots & \ddots & \vdots \\ (X_n-\mu_n)(X_1-\mu_1) & (X_n-\mu_n)(X_2-\mu_2) & \dots & (X_n-\mu_n)^2 \end{pmatrix}$$

$$= \begin{pmatrix} \mathrm{E}(X_1-\mu_1)^2 & \mathrm{E}(X_1-\mu_1)(X_2-\mu_2) & \dots & \mathrm{E}(X_1-\mu_1)(X_n-\mu_n) \\ \mathrm{E}(X_2-\mu_2)(X_1-\mu_1) & \mathrm{E}(X_2-\mu_2)^2 & \dots & \mathrm{E}(X_2-\mu_2)(X_n-\mu_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}(X_n-\mu_n)(X_1-\mu_1) & \mathrm{E}(X_n-\mu_n)(X_2-\mu_2) & \dots & \mathrm{E}(X_n-\mu_n)^2 \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_{nn} \end{pmatrix}.$$

Example

Find the covariance matrix for the two random variables X_1 and X_2 when their joint probability function $p_{12}(x_1, x_2)$ is represented by following table:

x_2 x_1	0	1	$p_1(x_1)$
-1	0.24	0.06	0.3
0	0.16	0.14	0.3
1	0.40	0.00	0.4
$p_2(x_1)$	0.8	0.2	1

Notice that: $\mu_1 = E(X_1) = 0.1$ and $\mu_2 = E(X_2) = 0.2$.

Example

$$\sigma_{11} = E(X_1 - \mu_1)^2 = \sum_{\text{all } x_1} (x_i - 0.1)^2 p_1(x_1)$$

$$= (-1 - .1)^2(.3) + (0 - .1)^2(.3) + (1 - .1)^2(.4) = 0.69$$

$$\sigma_{22} = E(X_2 - \mu_2)^2 = \sum_{\text{all } x_2} (x_i - 0.2)^2 p_2(x_2)$$

$$= (0 - .2)^2(.8) + (1 - .2)^2(.2) = .16$$

$$\sigma_{12} = E(X_1 - \mu_1)(X_2 - \mu_2)$$

$$= \sum_{\text{all pairs } (x_1, x_2)} (x_1 - 0.1)(x_2 - 0.2)p_{12}(x_1, x_2)$$

$$= (-1 - .1)(0 - .2)(.24) + (-1 - .1)(1 - .2)(.06) + \cdots = -0.08$$

$$\sigma_{21} = E(X_2 - \mu_2)(X_1 - \mu_1) = \sigma_{12} = -0.08.$$

Thus, the mean and covariance matrix of X are given, respectively, by:

$$\mu = E(X) = \begin{pmatrix} 0.1 \\ 0.2 \end{pmatrix} \text{ and } Var(X) = \Sigma = \begin{pmatrix} 0.69 & -0.08 \\ -0.08 & 0.16 \end{pmatrix}.$$

Multivariate Normal

The n random variables $X^T = (X_1, \ldots, X_n)$ have some probability density function (pdf) which is written as:

$$p(X) = p(X_1, \ldots, X_n).$$

This gives the likelihood of various combinations of X values. The most important multivariate pdf is the multivariate normal. It is specified in terms of its mean vector μ and its variance matrix Σ. The formula of the multivariate normal is given by:

$$\boxed{p(X) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(X-\mu)^T \Sigma^{-1}(X-\mu)\right).}$$

Compactly the latter is stated as:

$$\boxed{X \sim N(\mu, \Sigma).}$$

Properties

Let X and Y be random matrices of the same dimension and let A and B be conformable matrices of constants. Then,

- $E(X+Y) = E(X) + E(Y)$.
- $E(AXB) = A\,E(X)\,B$.
- $Var(AX) = A\,Var(X)\,A^T$.

Example

Let $X \in \Re^n$ have a positive definite covariance matrix Σ. Furthermore, let the Cholesky factor of $\Sigma = CC^T$. Find the covariance matrix of $Z = C^{-1}X$.

$$\begin{aligned} Var(Z) &= Var(C^{-1}X) = C^{-1}\,Var(X)\,C^{-T} \\ &= C^{-1}\Sigma C^{-T} = C^{-1}CC^TC^{-T} \\ &= I_n. \end{aligned}$$