Lecture 3

Normal Linear Models

1. Multi-dimensional data

2-dimensional data

No.	parent	Adult child
1	68.2	63.3
2	71.8	72.5
3	64.4	69.2
⋯	⋯	⋯
i	x_i	y_i
⋯	⋯	⋯
928	x_{928}	y_{928}

p-dimensional data

p variables

n samples

	x_1	x_2	$\cdots$	x_j	$\cdots$	x_p
1						
2						
$\vdots$						
i	x_{i1}	x_{i2}	$\cdots$	x_{ij}	$\cdots$	x_{ip}
$\vdots$						
n						

Matrix notation

ith data

$$\mathbf{X}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix}$$

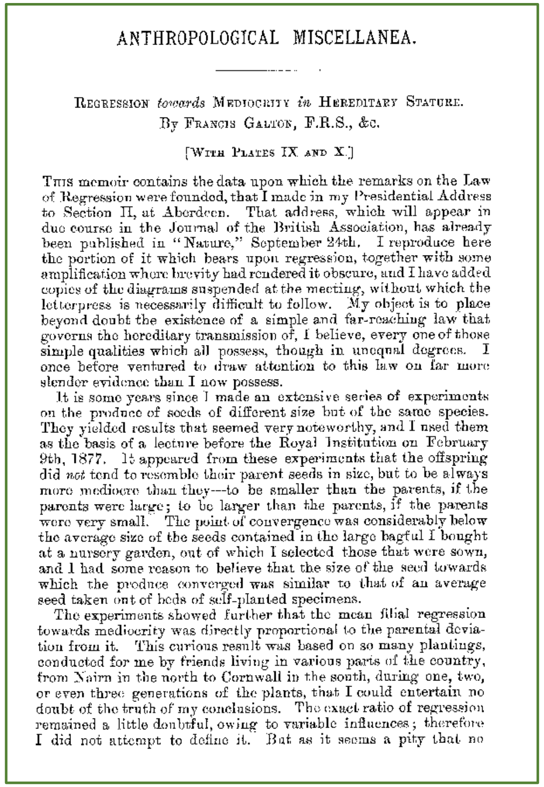
data matrix

$$\mathbf{D} = [\mathbf{X}_1 \; \mathbf{X}_2 \cdots \mathbf{X}_i \cdots \mathbf{X}_n]^T$$

1. Multi-dimensional data: Frequency table

	Mid-Heights of Parents (x)											
Heights of Adult Children (y)		below	64.5	65.5	66.5	67.5	68.5	69.5	70.5	71.5	72.5	above	sum
	above							5	3	2	4		14
	73.2						3	4	3	2	2	3	17
	72.2			1		4	4	11	4	9	7	1	41
	71.2			2		11	18	20	7	4	2		64
	70.2			5	4	19	21	25	14	10	1		99
	69.2	1	2	7	13	38	48	33	18	5	2		167
	68.2	1		7	14	28	34	20	12	3	1		120
	67.2	2	5	11	17	38	31	27	3	4			138
	66.2	2	5	11	17	36	25	17	1	3			117
	65.2	1	1	7	2	15	16	4	1	1			48
	64.2	4	4	5	5	14	11	16					59
	63.2	2	4	9	3	5	7	1	1				32
	62.2		1		3	3							7
	below	1	1	1			1		1				5
	sum	14	23	66	78	211	219	183	68	43	19	4	928

F. Galton:
Regression towards mediocrity in hereditary stature, Anthropological Miscellanea (1886)



ANTHROPOLOGICAL MISCELLANEA.

REGRESSION *towards* MEDIOCRITY *in* HEREDITARY STATURE.
By FRANCIS GALTON, F.R.S., &c.

[WITH PLATES IX AND X.]

THIS memoir contains the data upon which the remarks on the Law of Regression were founded, that I made in my Presidential Address to Section H, at Aberdeen. That address, which will appear in due course in the Journal of the British Association, has already been published in "Nature," September 24th. I reproduce here the portion of it which bears upon regression, together with some amplification where brevity had rendered it obscure, and I have added copies of the diagrams suspended at the meeting, without which the letterpress is necessarily difficult to follow. My object is to place beyond doubt the existence of a simple and far-reaching law that governs the hereditary transmission of, I believe, every one of those simple qualities which all possess, though in unequal degrees. I once before ventured to draw attention to this law on far more slender evidence than I now possess.

It is some years since I made an extensive series of experiments on the produce of seeds of different size but of the same species. They yielded results that seemed very noteworthy, and I used them as the basis of a lecture before the Royal Institution on February 9th, 1877. It appeared from these experiments that the offspring did *not* tend to resemble their parent seeds in size, but to be always more mediocre than they—to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were very small. The point of convergence was considerably below the average size of the seeds contained in the large bagful I bought at a nursery garden, out of which I selected those that were sown, and I had some reason to believe that the size of the seed towards which the produce converged was similar to that of an average seed taken out of beds of self-planted specimens.

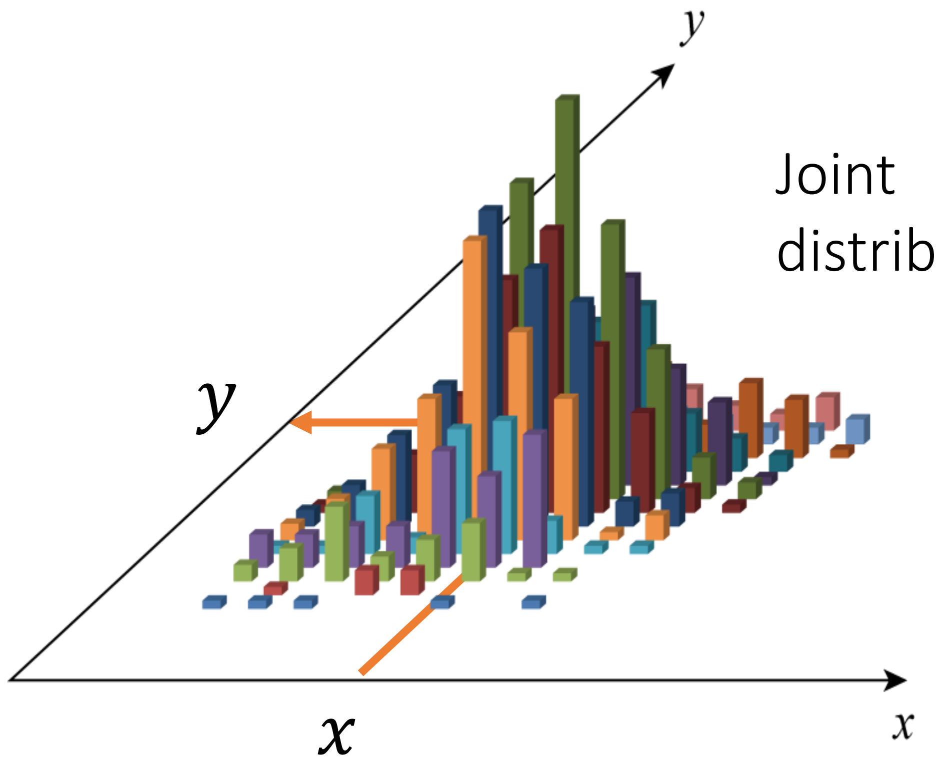
The experiments showed further that the mean filial regression towards mediocrity was directly proportional to the parental deviation from it. This curious result was based on so many plantings, conducted for me by friends living in various parts of the country, from Nairn in the north to Cornwall in the south, during one, two, or even three generations of the plants, that I could entertain no doubt of the truth of my conclusions. The exact ratio of regression remained a little doubtful, owing to variable influences; therefore I did not attempt to define it. But as it seems a pity that no

2. Regression analysis

y

Heights of Adult Children	Mid-heights of Parents											
	below	64.5	65.5	66.5	67.5	68.5	69.5	70.5	71.5	72.5	above	sum
above							5	3	2	4		14
73.2						3	4	3	2	2	3	17
72.2			1		4	4	11	4	9	7	1	41
71.2			2		11	18	20	7	4	2		64
70.2			5	4	19	21	25	14	10	1		99
69.2	1	2	7	13	38	48	33	18	5	2		167
68.2	1		7	14	28	34	20	12	3	1		120
67.2	2	5	11	17	38	31	27	3	4			138
66.2	2	5	11	17	36	25	17	1	3			117
65.2	1	1	7	2	15	16	4	1	1			48
64.2	4	4	5	5	14	11	16					59
63.2	2	4	9	3	5	7	1	1				32
62.2		1		3	3							7
below	1	1	1			1		1				5
sum	14	23	66	78	211	219	183	68	43	19	4	928

x



Joint distribution

We like to know a *reasonable formula*

$x \Rightarrow y$

3. Models for response

input
explanatory variables

- one-variable
  x
- two-variable
  $\boldsymbol{x} = [x_1 \ \ x_2]$
- or more
  $\boldsymbol{x} = [x_1 \ \cdots \ x_p]$

outcome
response

deterministic system → y

$$y = f(x)$$
$$y = f(\boldsymbol{x}) = f(x_1, x_2)$$

deterministic linear system → y

$$y = \beta x \quad (\text{ or } \quad y = \beta x + \gamma \ )$$
$$y = x_1\beta_1 + x_2\beta_2$$

stochastic linear system → y

$$Y = \beta x + e$$
$$Y = x_1\beta_1 + x_2\beta_2 + e$$
$$e \sim N(0, \sigma^2)$$

5. Normal linear models

i-th input ($i = 1,2,\cdots,n$)

- one-variable x_i
- p-variable $\boldsymbol{x}_i = [x_{i1} \ \cdots \ x_{ip}]$

$\Rightarrow$ stochastic linear system $\Rightarrow$ outcome y_i

We consider y_i is a realized value of a random variable Y_i where

$$Y_i = \beta x_i + e_i \qquad \text{or} \qquad Y_i = x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + e_i$$

$\{e_1, e_2, \cdots, e_n\}$ are independent random variables obeying $N(0, \sigma^2)$

In matrix expression:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \quad \text{where} \quad \boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_i \\ \vdots \\ e_n \end{bmatrix}$$

5. Normal linear models: An example

x_i (Parents’ height) $\Rightarrow$ stochastic linear system $\Rightarrow$ y_i (Child’s height)

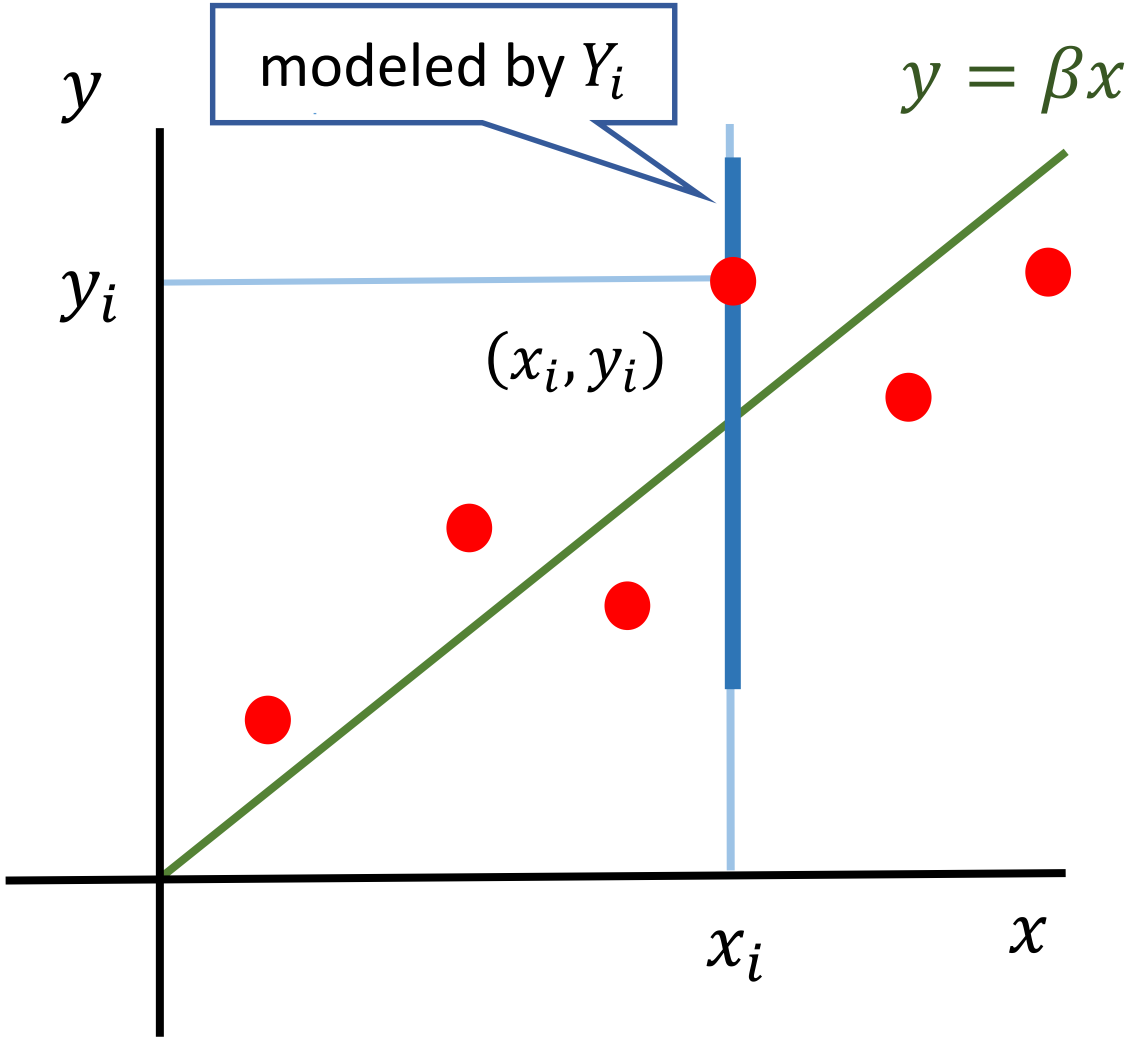
$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{bmatrix}$, $\boldsymbol{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{bmatrix}$, $\boldsymbol{\beta} = [\beta]$, $\boldsymbol{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_i \\ \vdots \\ e_n \end{bmatrix}$

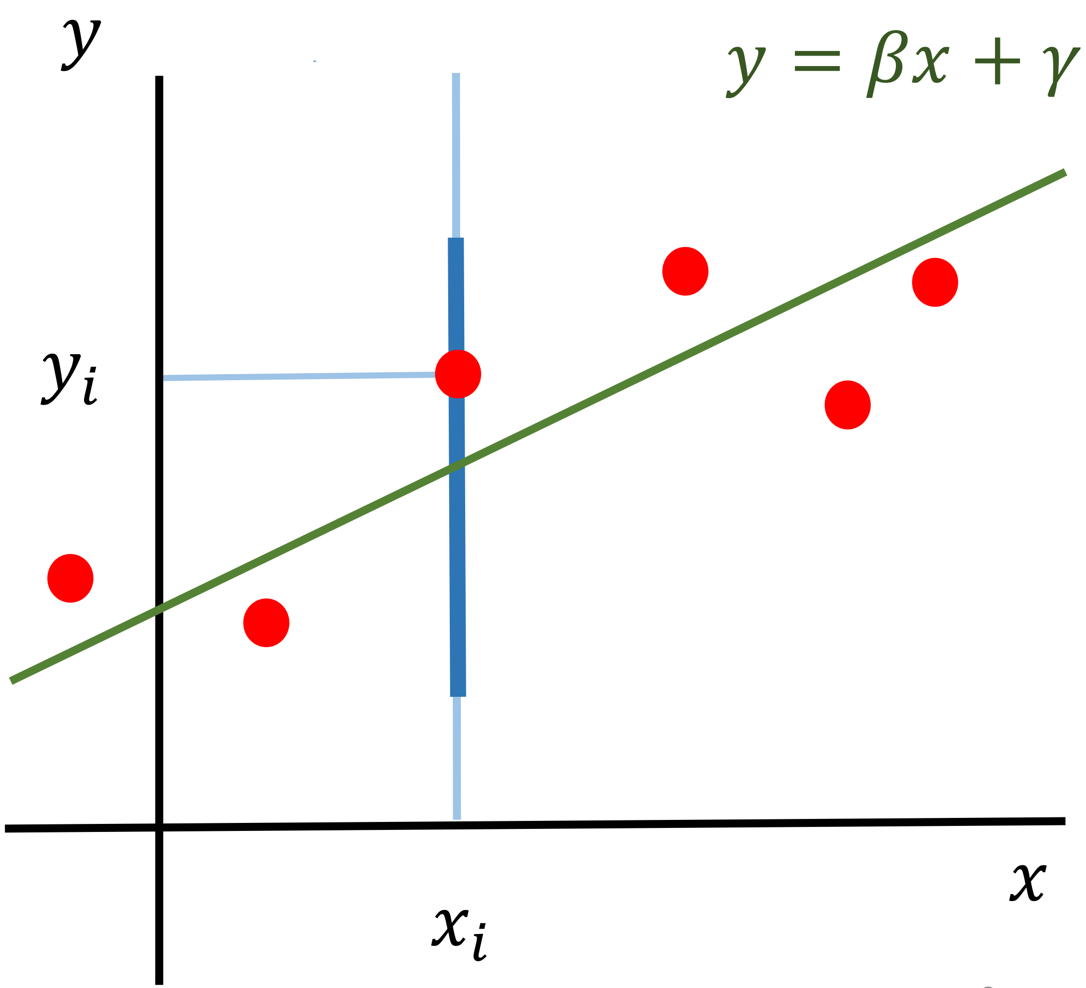
where $\{e_1, e_2, \cdots, e_n\}$ are iid random variables obeying $N(0, \sigma^2)$

Basic Problem: Given observations $y_1, y_2, \cdots, y_n$ we determine the best β

6. Scatter plot

Problem: Find the best parameters β, γ





7. Maximum likelihood estimation

How to determine the best $\boldsymbol{\beta}$ in the normal linear model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$

1-dimensional input (case of single explanatory variable)

$Y_i = \beta x_i + e_i \quad i = 1,2,\cdots,n$ (size of data)

$\{e_1, e_2, \cdots, e_n\}$ are iid random variables obeying $N(0, \sigma^2)$

Then $Y_i \sim N(\mu_i, \sigma^2)$ with $\mu_i = \beta x_i$ so the probability density function is given by

$$f_{Y_i}(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \mu_i)^2}{2\sigma^2}\right\} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \beta x_i)^2}{2\sigma^2}\right\}$$

The log-likelihood function

$$\log L = \log \prod_{i=1}^{n} f_{Y_i}(y_i) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta x_i)^2$$

The log-likelihood function

$$\log L = \log \prod_{i=1}^{n} f_{Y_i}(y_i) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta x_i)^2$$

We need to find β which maximizes $\log L$

$$\frac{\partial}{\partial\beta}\log L = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}2(y_i - \beta x_i)(-x_i) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta x_i)x_i = 0$$

$$\Rightarrow \quad \sum_{i=1}^{n} x_i y_i - \beta \sum_{i=1}^{n} x_i^2 = 0$$

$$\Rightarrow \quad \beta = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Check the details!

In matrix form

$$\beta = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \qquad \boldsymbol{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{bmatrix}, \qquad \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}$$

p-dimensional input (case of p explanatory variables)

Recall the case of 1-dimensional input

$$Y_i = \beta x_i + e_i \quad i = 1,2,\cdots,n \text{ (size of data)}$$

$\{e_1, e_2, \cdots, e_n\}$ are independent random variables obeying $N(0, \sigma^2)$

p-dimensional input

$$Y_i = \sum_{j=1}^{p} x_{ij}\beta_j + e_i \qquad (i = 1,2,\cdots,n) \qquad \{e_1, e_2, \cdots, e_n\} \text{ iid obeying } N(0, \sigma^2)$$

Then $Y_i \sim N(\mu_i, \sigma^2)$ with $\mu_i = \sum_{j=1}^{p} x_{ij}\beta_j$

so the probability density function is given by

$$f_{Y_i}(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \mu_i)^2}{2\sigma^2}\right\}$$

The log-likelihood function

$$\log L = \log \prod_{i=1}^{n} f_{Y_i}(y_i) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2 \qquad \mu_i = \sum_{j=1}^{p} x_{ij}\beta_j$$

We need to find β_k which maximize $\log L$

$$\frac{\partial}{\partial \beta_k}\log L = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(y_i - \mu_i)\left(-\frac{\partial \mu_i}{\partial \beta_k}\right) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)x_{ik} = 0$$

$$\Rightarrow \quad \sum_{i=1}^{n} y_i x_{ik} - \sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}\beta_j x_{ik} = 0 \qquad (k = 1,2,\cdots,p)$$

$\Rightarrow$ This is a linear system and is solvable. $\Rightarrow$ Matrix notation is essential!

Matrix notation

$$\sum_{i=1}^{n} y_i x_{ik} - \sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}\beta_j x_{ik} = 0 \qquad (k = 1,2,\cdots,p) \quad \Leftrightarrow \quad (\boldsymbol{X}^T\boldsymbol{y})_k - (\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta})_k = 0$$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_p \end{bmatrix}$$

$$\Leftrightarrow \quad \boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}$$

$$\Leftrightarrow \quad \boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Check the details!

Theorem. For a normal linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$$

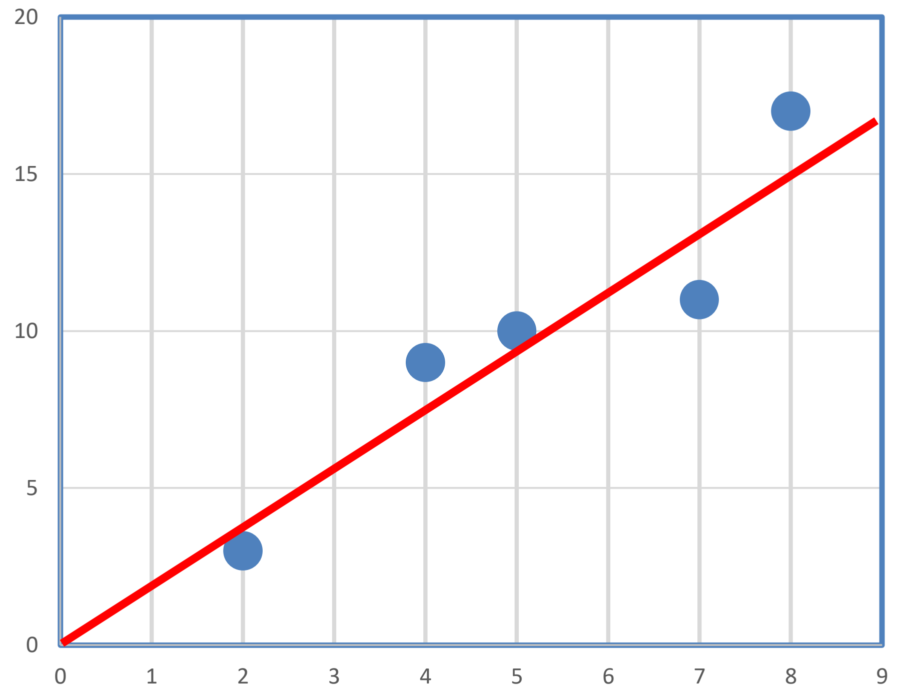
the maximum likelihood estimation is given by

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

8. Example

Based on a physical principle two variables x and y are known to be related as $y = \beta x$. In an experiment x is controlled very precisely, and we measure y. The result of five experiments is as follows. Determine β.

x	2	4	5	7	8
y	3	9	10	11	17

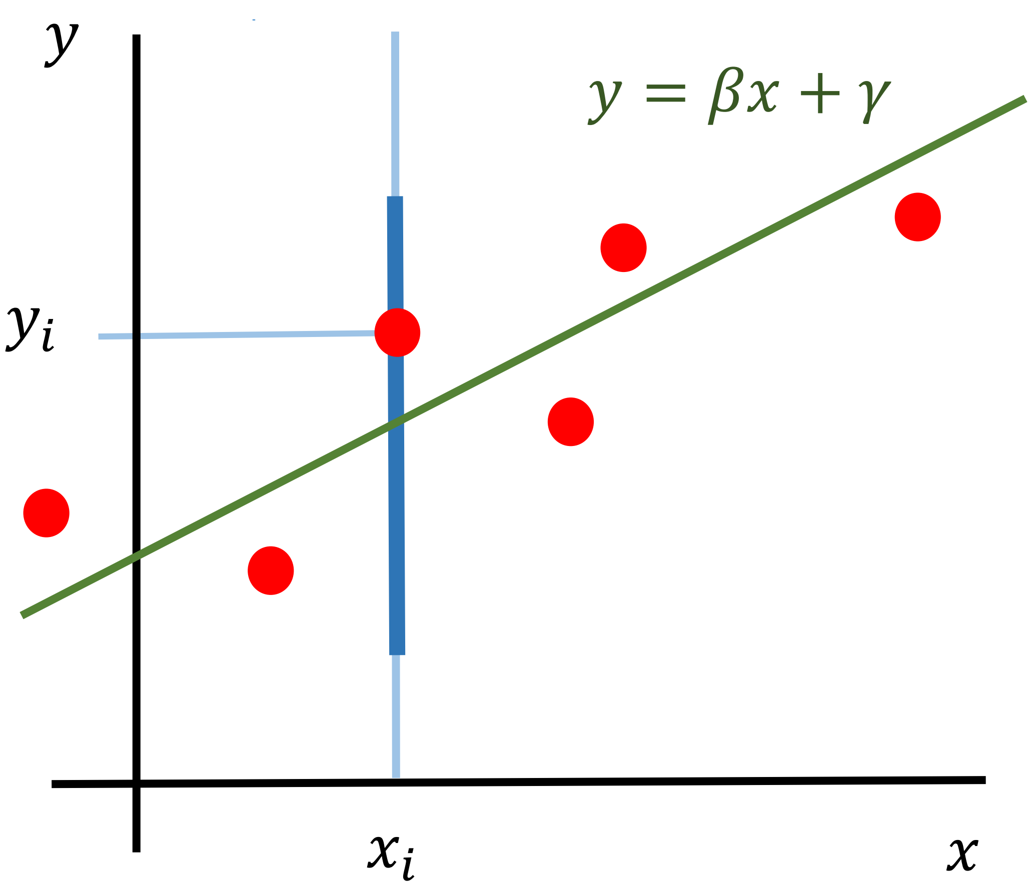


We apply the formula $\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}^T\boldsymbol{y}$ where

$$\boldsymbol{\beta} = [\beta], \quad \boldsymbol{X} = \begin{bmatrix} 2 \\ 4 \\ 5 \\ 7 \\ 8 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} 3 \\ 9 \\ 10 \\ 11 \\ 17 \end{bmatrix} \quad \Rightarrow \quad \boldsymbol{X}^T\boldsymbol{X} = [158] \quad \boldsymbol{X}^T\boldsymbol{y} = [305] \quad \beta = \frac{305}{158} = 1.93$$

9. Regression lines

Ask for the best parameters β, γ



Setting

dummy variable

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} x_1 & 1 \\ \vdots & 1 \\ x_i & 1 \\ \vdots & 1 \\ x_n & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta \\ \gamma \end{bmatrix}$$

Our normal linear model is given by

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e} \Leftrightarrow y_i = \beta x_i + \gamma + e_i$$

and the maximum likelihood estimation is given by

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

9. Regression lines: An explict form

The maximum likelihood estimation:

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

where $\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}$, $\boldsymbol{X} = \begin{bmatrix} x_1 & 1 \\ \vdots & 1 \\ x_i & 1 \\ \vdots & 1 \\ x_n & 1 \end{bmatrix}$, $\boldsymbol{\beta} = \begin{bmatrix} \beta \\ \gamma \end{bmatrix}$

First we note that

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{bmatrix} \qquad \boldsymbol{X}^T\boldsymbol{y} = \begin{bmatrix} \sum x_i y_i \\ \sum y_i \end{bmatrix}$$

It is convenient to introduce

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\,,$$

$$\sigma_{xx} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 \qquad \sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$$

Then we have

$$\boldsymbol{X}^T\boldsymbol{X} = n\begin{bmatrix} \sigma_{xx} + \bar{x}^2 & \bar{x} \\ \bar{x} & 1 \end{bmatrix} \qquad \boldsymbol{X}^T\boldsymbol{y} = n\begin{bmatrix} \sigma_{xy} + \bar{x}\bar{y} \\ \bar{y} \end{bmatrix}$$

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

$$= \frac{1}{\sigma_{xx}}\begin{bmatrix} 1 & -\bar{x} \\ -\bar{x} & \sigma_{xx} + \bar{x}^2 \end{bmatrix}\begin{bmatrix} \sigma_{xy} + \bar{x}\bar{y} \\ \bar{y} \end{bmatrix}$$

$$= \frac{1}{\sigma_{xx}}\begin{bmatrix} \sigma_{xy} \\ -\bar{x}\sigma_{xy} + \bar{y}\sigma_{xx} \end{bmatrix}$$

Namely,

$$\boldsymbol{\beta} = \begin{bmatrix} \beta \\ \gamma \end{bmatrix} = \frac{1}{\sigma_{xx}} \begin{bmatrix} \sigma_{xy} \\ -\bar{x}\sigma_{xy} + \bar{y}\sigma_{xx} \end{bmatrix}$$

Thus,

$$y = \beta x + \gamma = \frac{\sigma_{xy}}{\sigma_{xx}} x - \frac{\sigma_{xy}}{\sigma_{xx}} \bar{x} + \bar{y}$$

$$\Longleftrightarrow \quad y - \bar{y} = \frac{\sigma_{xy}}{\sigma_{xx}} (x - \bar{x})$$

$$\Longleftrightarrow \quad \frac{y - \bar{y}}{\sqrt{\sigma_{yy}}} = \frac{\sigma_{xy}}{\sqrt{\sigma_{xx}\sigma_{yy}}} \frac{x - \bar{x}}{\sqrt{\sigma_{xx}}}$$

Correlation coefficient

$$\rho_{xy} = \frac{\sigma_{xy}}{\sqrt{\sigma_{xx}\sigma_{yy}}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Theorem Regression line of y on x is given by

$$\frac{y - \bar{y}}{\sigma_y} = \rho_{xy} \frac{x - \bar{x}}{\sigma_x}$$

Note: the roles of x and y are not symmetric. Regression line of x on y is given by

$$\rho_{xy} \frac{y - \bar{y}}{\sigma_y} = \frac{x - \bar{x}}{\sigma_x}$$

Exercise 5 (10min)

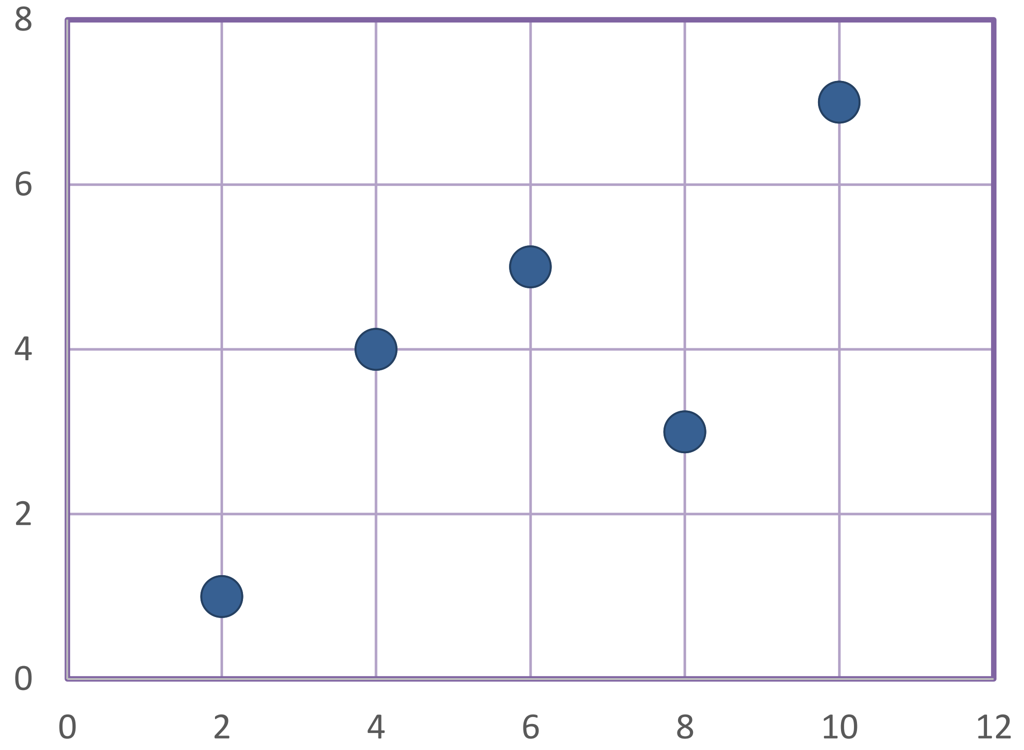
We have the following data.

x	6	2	10	4	8
y	5	1	7	4	3

Find the regression line of y on x,

by using the formula

$$\frac{y-\bar{y}}{\sigma_y} = \rho_{xy}\,\frac{x-\bar{x}}{\sigma_x}$$



[Hint] completing the following table to obtain $\bar{x}, \bar{y}, \sigma_x^2, \sigma_y^2, \sigma_{xy}$ and ρ_{xy}

						sum	mean
x	6	2	10	4	8		
y	5	1	7	4	3		
x^2							
y^2							
xy							

10. Least squares estimation

Normal linear model

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$$

$$\Leftrightarrow \quad Y_i = \sum_{k=1}^{p} x_{ik}\beta_k + e_i$$

mean vector

$$\mathbf{E}[\boldsymbol{Y}] = \boldsymbol{X\beta} + \mathbf{E}[\boldsymbol{e}] = \boldsymbol{X\beta}$$

$$\Leftrightarrow \ \mathbf{E}[Y_i] = (\boldsymbol{X\beta})_i = \sum_{k=1}^{p} x_{ik}\beta_k$$

variance-covariance matrix

$$\mathbf{V}[\boldsymbol{Y}] = \mathbf{E}[(\boldsymbol{Y} - \boldsymbol{X\beta})(\boldsymbol{Y} - \boldsymbol{X\beta})^T]$$

$$\begin{aligned}
\Leftrightarrow \ (\mathbf{V}[\boldsymbol{Y}])_{ij} &= \mathbf{E}\left[\left((\boldsymbol{Y} - \boldsymbol{X\beta})(\boldsymbol{Y} - \boldsymbol{X\beta})^T\right)_{ij}\right] \\
&= \mathbf{E}\left[(\boldsymbol{Y} - \boldsymbol{X\beta})_i(\boldsymbol{Y} - \boldsymbol{X\beta})_j\right] \\
&= \mathbf{E}\left[(Y_i - \mathbf{E}[Y_i])\left(Y_j - \mathbf{E}[Y_j]\right)\right] \\
&= \mathbf{Cov}\left(Y_i, Y_j\right) = \sigma^2\delta_{ij}
\end{aligned}$$

Because $\{e_1, e_2, \cdots, e_n\}$ are independent random variables obeying $N(0, \sigma^2)$ by assumption.

Suppose we are given data $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_n, y_n)$

Consider

$$S(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X\beta})^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X\beta})$$

where

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{V} = \begin{bmatrix} \sigma^2 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & \sigma^2 & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & \cdots & \sigma^2 \end{bmatrix}$$

$$S(\boldsymbol{\beta}) = \sigma^{-2} \sum_{i=1}^{n} (y_i - (\boldsymbol{X\beta})_i)^2 = \sigma^{-2} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2$$

Minimize $$S(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X\beta})^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X\beta}) = \sigma^{-2} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

$$\frac{\partial}{\partial \beta_k} S(\boldsymbol{\beta}) = \sigma^{-2} \sum_{i=1}^{n} 2 \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right) (-x_{ik}) = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} y_i x_{ik} = \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij} x_{ik} \beta_j$$

$$\Leftrightarrow (\boldsymbol{X}^T \boldsymbol{y})_k = \sum_{j=1}^{p} (\boldsymbol{X}^T \boldsymbol{X})_{kj} \beta_j = (\boldsymbol{X}^T \boldsymbol{X\beta})_k$$

$$\Leftrightarrow \boldsymbol{X}^T \boldsymbol{y} = \boldsymbol{X}^T \boldsymbol{X\beta}$$

$$\Leftrightarrow \boldsymbol{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

Check the details!

Theorem For a normal linear model

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$$

the least squares estimation is given by

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

Note: The result is the same as the one obtained by maximum likelihood estimation.

11. Two-dimensional normal distribution $N(\boldsymbol{m}, \Sigma)$

- mean vector $\quad \boldsymbol{m} = \begin{bmatrix} a \\ b \end{bmatrix}$
- variance-covariance matrix $\quad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$

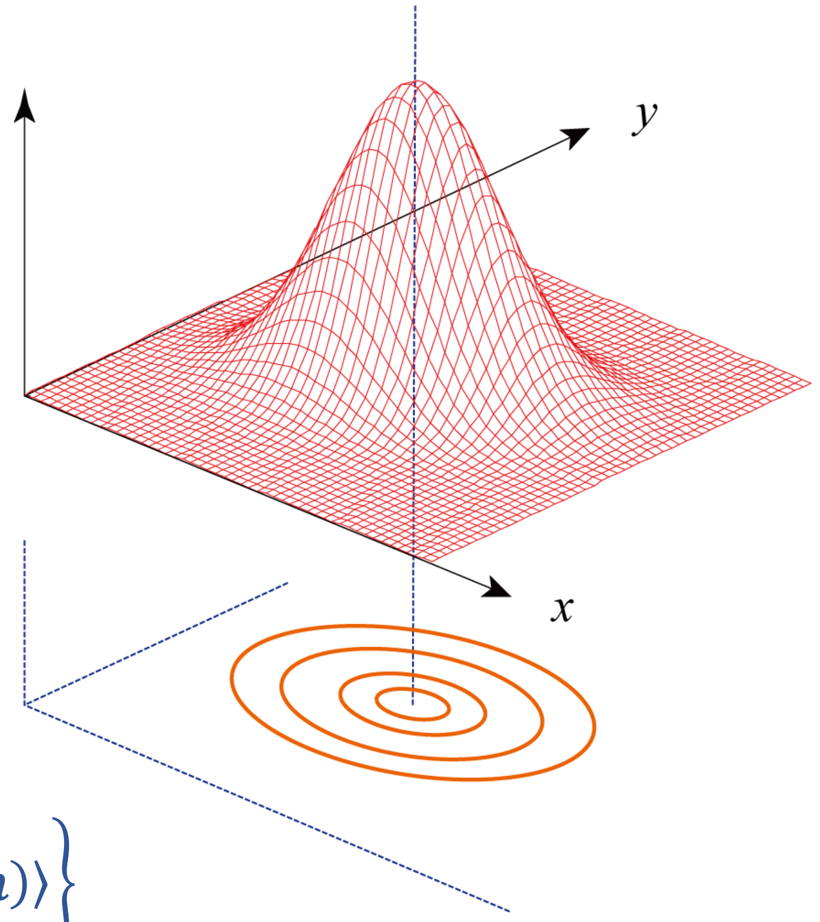
  necessarily symmetric

  $$\sigma_{12} = \sigma_{21}$$

  and positive definite

  $\langle \boldsymbol{x}, A\boldsymbol{x} \rangle > 0$ for all $\boldsymbol{x}$ with $\boldsymbol{x} \neq \boldsymbol{0}$.

  Then $\quad |\Sigma| = \det \Sigma \neq 0$



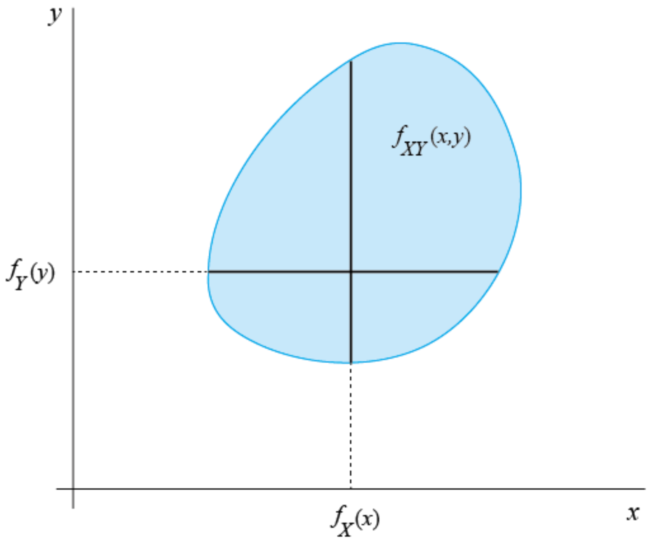
- density function of $N(\boldsymbol{m}, \Sigma)$

$$f(x, y) = f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^2 |\Sigma|}} \exp\left\{ -\frac{1}{2} \langle (\boldsymbol{x} - \boldsymbol{m}), \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{m}) \rangle \right\}$$

Note: $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^T \boldsymbol{y}$ (LHS is a matrix form)

11. Two-dimensional normal distribution $N(\boldsymbol{m}, \Sigma)$

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^2|\Sigma|}} \exp\left\{-\frac{1}{2}\langle(\boldsymbol{x}-\boldsymbol{m}), \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{m})\rangle\right\} = f_{XY}(x,y)$$



Marginal distributions

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y)\,dy \qquad f_Y(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y)\,dx$$

$$m_X = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xf(x,y)dxdy = a \qquad m_Y = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} yf(x,y)dxdy = b$$

$$\sigma_X^2 = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (x-a)^2 f(x,y)dxdy = \sigma_{11} \qquad \sigma_Y^2 = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (y-b)^2 f(x,y)dxdy = \sigma_{22}$$

$$\sigma_{XY} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (x-a)(y-b)f(x,y)dxdy = \sigma_{12} = \sigma_{21}$$

Check the details!

THEOREM: Assume that a random vector (X, Y) obeys 2-dimensional normal distribution $N(\boldsymbol{m}, \Sigma)$. Then $\mathbf{Cov}(X, Y) = 0$ implies that X and Y are independent.

PROOF. Note first that

$$\Sigma = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix} = \begin{bmatrix} \sigma_{xx} & 0 \\ 0 & \sigma_{yy} \end{bmatrix}$$

$$\sigma_{xx} = \sigma_x^2, \quad \sigma_{yy} = \sigma_y^2,$$

$$\sigma_{xy} = \sigma_{yx} = \mathbf{Cov}(X, Y)$$

Then

$$\langle (\boldsymbol{x} - \boldsymbol{m}), \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{m}) \rangle$$

$$= \left\langle \begin{bmatrix} x - m_X \\ y - m_Y \end{bmatrix}, \begin{bmatrix} \sigma_{xx}^{-1} & 0 \\ 0 & \sigma_{yy}^{-1} \end{bmatrix} \begin{bmatrix} x - m_X \\ y - m_Y \end{bmatrix} \right\rangle$$

$$= \sigma_{xx}^{-1}(x - m_X)^2 + \sigma_{xx}^{-1}(x - m_X)^2$$

$$|\Sigma| = \sigma_{xx}\sigma_{yy}$$

$$f(x, y) = \frac{1}{\sqrt{(2\pi)^2 |\Sigma|}} \exp\left\{ -\frac{1}{2} \langle (\boldsymbol{x} - \boldsymbol{m}), \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{m}) \rangle \right\}$$

$$= \frac{1}{\sqrt{(2\pi)^2 \sigma_{xx}\sigma_{yy}}} \exp\left\{ -\frac{1}{2} \{ \sigma_{xx}^{-1}(x - m_X)^2 + \sigma_{yy}^{-1}(y - m \quad)^2 \} \right\}$$

$$= \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left\{ -\frac{(x - m_X)^2}{2\sigma_x^2} \right\} \times \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left\{ -\frac{(y - m_Y)^2}{2\sigma_y^2} \right\}$$

$$= f_X(x) f_Y(y)$$

Since the joint density function is factorized, X and Y are independent.

Check the details!

Problem 9

Assume that a random vector (X, Y) obeys 2-dimensional normal distribution $N(\boldsymbol{m}, \Sigma)$, that is, the joint distribution is given by

$$f_{XY}(x, y) = \frac{1}{\sqrt{(2\pi)^2|\Sigma|}} \exp\left\{-\frac{1}{2}\langle(\boldsymbol{x}-\boldsymbol{m}), \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{m})\rangle\right\} \qquad \boldsymbol{m} = \begin{bmatrix} a \\ b \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

(1) Show by direct computation of the integral that the marginal distribution $f_X(x)$ obeys $N(a, \sigma_{11})$.

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x, y)\, dy$$

(2) Calculate the conditional expectation:

$$\mathbf{E}[Y|X = x] = \int_{-\infty}^{+\infty} y\, f_{Y|X}(y|x) dy \qquad \text{where} \quad f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}$$

(3) Examine that $y = \mathbf{E}[Y|X = x]$ is the regression line of y on x.

Problem 10

The right table shows 892 samples of (x, y), where x is the mid-height of parents and y the height of their child [Galton 1886].

(1) Find the correlation coefficient.

(2) Find the regression line of y on x.

	Mid-height parents (x)									
Adult Children (y)		64.5	65.5	66.5	67.5	68.5	69.5	70.5	71.5	72.5	sum
	73.2					3	4	3	2	2	14
	72.2		1		4	4	11	4	9	7	40
	71.2		2		11	18	20	7	4	2	64
	70.2		5	4	19	21	25	14	10	1	99
	69.2	2	7	13	38	48	33	18	5	2	166
	68.2		7	14	28	34	20	12	3	1	119
	67.2	5	11	17	38	31	27	3	4		136
	66.2	5	11	17	36	25	17	1	3		115
	65.2	1	7	2	15	16	4	1	1		47
	64.2	4	5	5	14	11	16				55
	63.2	4	9	3	5	7	1	1			30
	62.2	1		3	3						7
	sum	22	65	78	211	218	178	64	41	15	892