STOR 455
STATISTICAL METHODS I

Jan Hannig

Matrix operations - review

- STAPLE YOUR HOMEWORKS!
- Multiplication
  - Example on the board
  - Multiplication of matrices has a geometric explanation.

Inverse Matrix

- $\boldsymbol{A}^{-1}$ is a matrix such that $\boldsymbol{A}.\,\boldsymbol{A}^{-1}=\boldsymbol{A}^{-1}.\boldsymbol{A}=\boldsymbol{I}$

$$A = \begin{bmatrix} 7 & 2 \\ 10 & 3 \end{bmatrix} \qquad A^{-1} = \begin{bmatrix} 3 & -2 \\ -10 & 7 \end{bmatrix}$$

- Easy to find for *2×2*

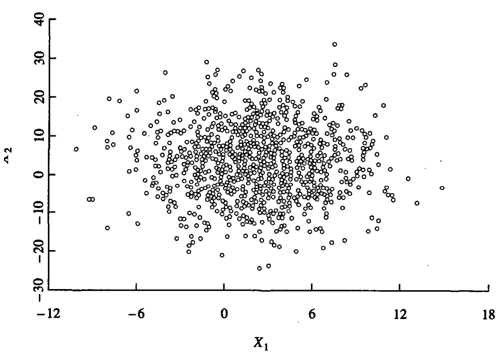
$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \qquad A^{-1} = \begin{bmatrix} a_{22}/d & -a_{12}/d \\ -a_{21}/d & a_{11}/d \end{bmatrix}$$

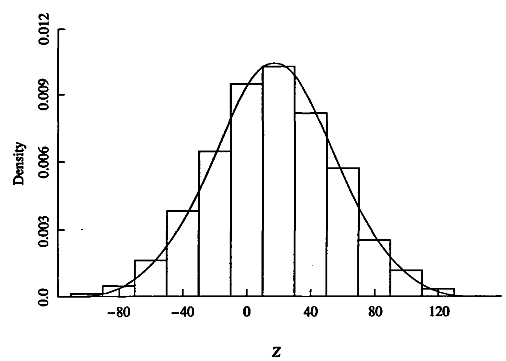
  – *Here* $d=\det(\boldsymbol{A})=a_{11}a_{22}-a_{12}a_{21}$

- Used for solving linear equation (see example on the board)

Multivariate Gaussian Distribution

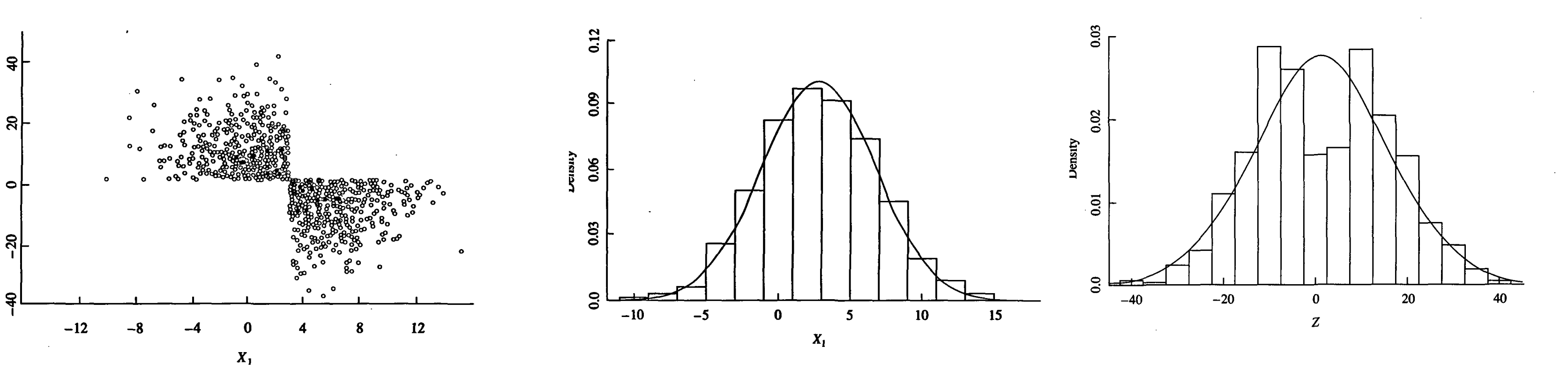
- Consider *random vector* $\boldsymbol{X}=(X_1,\ldots,X_d)'$
  – Each X_i is a random variable
- $\boldsymbol{X}$ is Gaussian if for all $a_1,\ldots,a_d$
  $Z=a_1X_1+\ldots+a_dX_d$ is Gaussian
  – Scatter plot is *elliptic,* histogram *is bell shaped*





Multivariate Gaussian

- It is not enough to be Gaussian for some ***a***. It needs to be for all.



Regression

- Chapter 2 has very good “philosophical” motivation of Regression.
  - We will only hit highlights – read the text!
- The main reasons for doing regression
  - Prediction – data based “guess” for unknown or future values
  - Model description – understanding the “underlying science”

Regression Population

- I this class we have **one** *response* (dependent) variable Y and one or more *predictor* (independent) variable $X_1,\ldots,X_p$
- Important idea
  - Y is expensive or impossible to measure (future)
  - $\boldsymbol{X}$ is easier to obtain or in our control (investment strategy, current values,…)
  - We have some observations of Y *and* $\boldsymbol{X}$ available to build a statistical model.

Regression Populations

- Examples
  - X – miles driven per year
    Y – cost of maintenance
  - X – age and weight of an individual
    Y – blood pressure
  - X – various variables measuring childhood development
    Y – ability to cope with stress
  - GIVE ME EXAMPLES

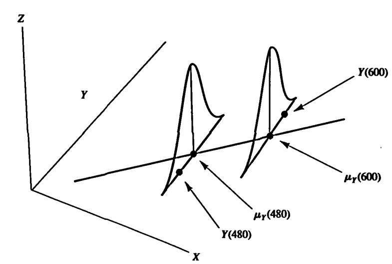
Regression population

A *Schematic Representation of a Trivariate Population with Response Variable* Y and Predictor Variables X_1 and X_2

Item Number I	Response Variable Y	Predictor Variable 1 (Explanatory Variable 1) X_1	Predictor Variable 2 (Explanatory Variable 2) X_2
1	Y_1	X_{11}	X_{12}
2	Y_2	X_{21}	X_{22}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	Y_I	X_{I1}	X_{I2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	Y_N	X_{N1}	X_{N2}

Goal of regression

- Based on data available find a prediction function $P_Y(x_1,\ldots,x_p)$ that best "fits" the data.
- Typically we will use "least square fit"
  - Select a function $P_Y(x_1,\ldots,x_p)$ so that $\Sigma(Y-P_Y(x_1,\ldots,x_p))^2$ is *minimized.*
  - Read the section on subpopulation – all possible values of *Y* with ***X*** held fixed.
    - $P_Y(x_1,\ldots,x_p) = \mu_Y(x_1,\ldots,x_p)$
    - $\sigma_Y(x_1,\ldots,x_p)$ is also important



Linear Regression

- Simple Linear Regression (straight line regression)
  - $\mu_Y(x) = \beta_0 + \beta_1 x$, $\sigma_Y(x) = \sigma$
  - Very useful in practice – we will start with this
- Multiple linear regression
  - More than one predictor
  - $\mu_Y(x_1, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$, $\sigma_Y(x) = \sigma$
  - We will cover this later in the year

Linear vs non-linear

- It is important that the linear regression function is linear in the *parameters*

$$
\begin{aligned}
\mu_Y(x) &= \beta_0 \\
\mu_Y(x) &= \beta_0 + \beta_1 x \\
\mu_Y(x) &= \beta_0 + \beta_1 x + \beta_2 x^2 \\
\mu_Y(x_1, x_2, x_3) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\
\mu_Y(x_1) &= \beta_0 + \beta_1 x_1^2 + \beta_2 x_1^{3/2} + \beta_3 / \ln|x_1| \\
\mu_Y(x_1, x_2) &= \beta_0 + \beta_1 e^{x_1} + \beta_2 x_2 + \beta_3 e^{x_1 x_2} \\
\mu_Y(x_1, x_2, x_3) &= \beta_0 + \beta_1 e^{-2x_1} + \beta_2 \sin(x_1 x_2) + \beta_3 x_1 \ln(x_2^2) \tan(x_3) \\
\mu_Y(x_1, x_2, x_3) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_1 x_3^2
\end{aligned}
$$

Linear vs non-linear

- Example of non-linear regression functions

$$
\begin{aligned}
\mu_Y(x_1) &= \beta_1 e^{\beta_2 x_1} \\
\mu_Y(x_1) &= \beta_0 + \beta_1 e^{\beta_2 x_1} \\
\mu_Y(x_1, x_2) &= \beta_0 + \beta_1 e^{\beta_2 x_1} + \beta_3 e^{\beta_4 x_2} \\
\mu_Y(x_1, x_2, x_3) &= \beta_0 x_1^{\beta_1} x_2^{\beta_2} x_3^{\beta_3} \\
\mu_Y(x_1, x_2) &= \beta_1 x_1 / (\beta_2 e^{\beta_3 x_2})
\end{aligned}
$$

SAS Example (Task 2.3.1)

- ```
  data car;
  infile 'T:\...\CAR.DAT';
  input carno mtcost price miles;
  run;
  ```
- ```
  proc contents data=car;
  run;
  ```
- ```
  data subpop;
  set car;
  if miles=14000;
  proc print data=subpop;
  run;
  ```
- ```
  proc chart data=car;
  hbar mtcost;
  run;
  ```
- ```
  proc plot data=car;
  plot mtcost*miles='*'/hpos=50 vpos=15;
  run;
  ```
  ```