Multivariate Analysis

CS 215 Fall 2024

Multivariate data

- Data reduction or summarization
  - Studying data as simply as possible without sacrificing useful information

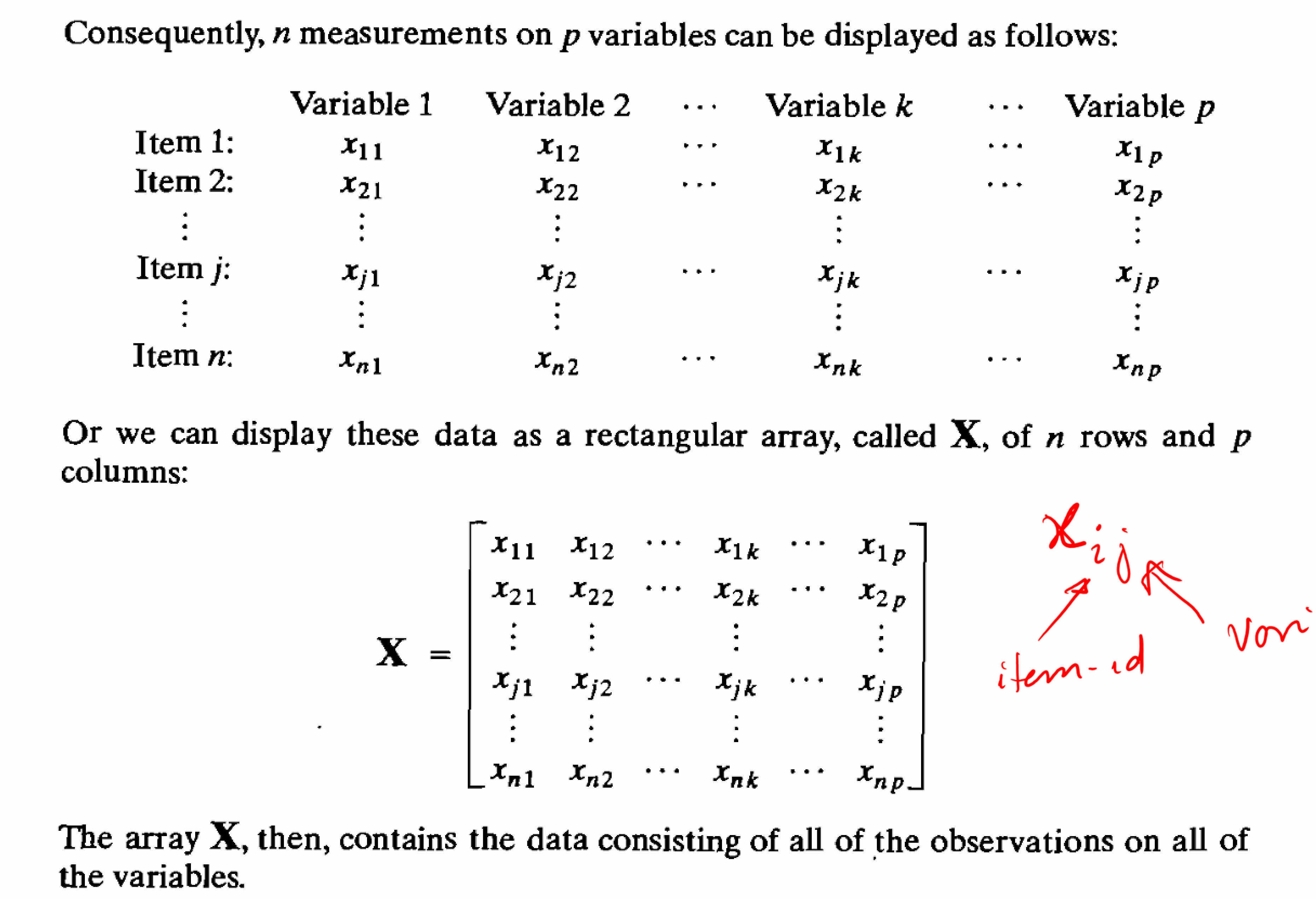
- Sorting and grouping
  - Clustering similar object together

- Investigating dependence among variables
  - Mutual independence, conditional independence etc. (already done)

- Prediction
  - Regression (already done)

- Hypothesis testing
  - Validate assumptions

Multivariate data organization.



Consequently, n measurements on p variables can be displayed as follows:

	Variable 1	Variable 2	$\cdots$	Variable k	$\cdots$	Variable p
Item 1:	x_{11}	x_{12}	$\cdots$	x_{1k}	$\cdots$	x_{1p}
Item 2:	x_{21}	x_{22}	$\cdots$	x_{2k}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
Item j:	x_{j1}	x_{j2}	$\cdots$	x_{jk}	$\cdots$	x_{jp}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
Item n:	x_{n1}	x_{n2}	$\cdots$	x_{nk}	$\cdots$	x_{np}

Or we can display these data as a rectangular array, called $\mathbf{X}$, of n rows and p columns:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}$$

$$\underset{\text{item-id}}{x_{i}}\,\underset{\text{variable-id}}{{}_{j}}$$

The array $\mathbf{X}$, then, contains the data consisting of all of the observations on all of the variables.

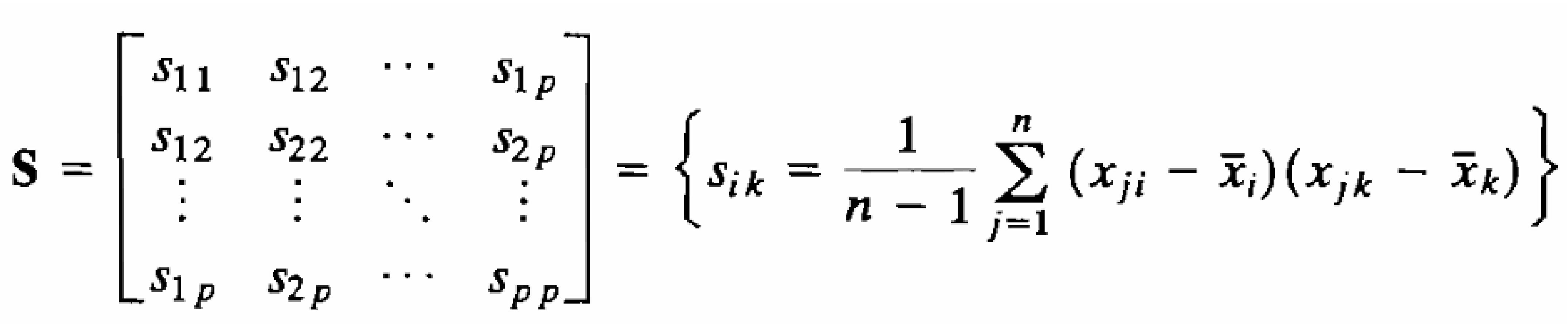
Multivariate descriptive statistics

Sample means
$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} \qquad \bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk} \qquad k = 1, 2, \ldots, p$$

Sample variances and covariances
$$\mathbf{S}_n = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} \qquad s_{ik} = \frac{1}{n}\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)$$

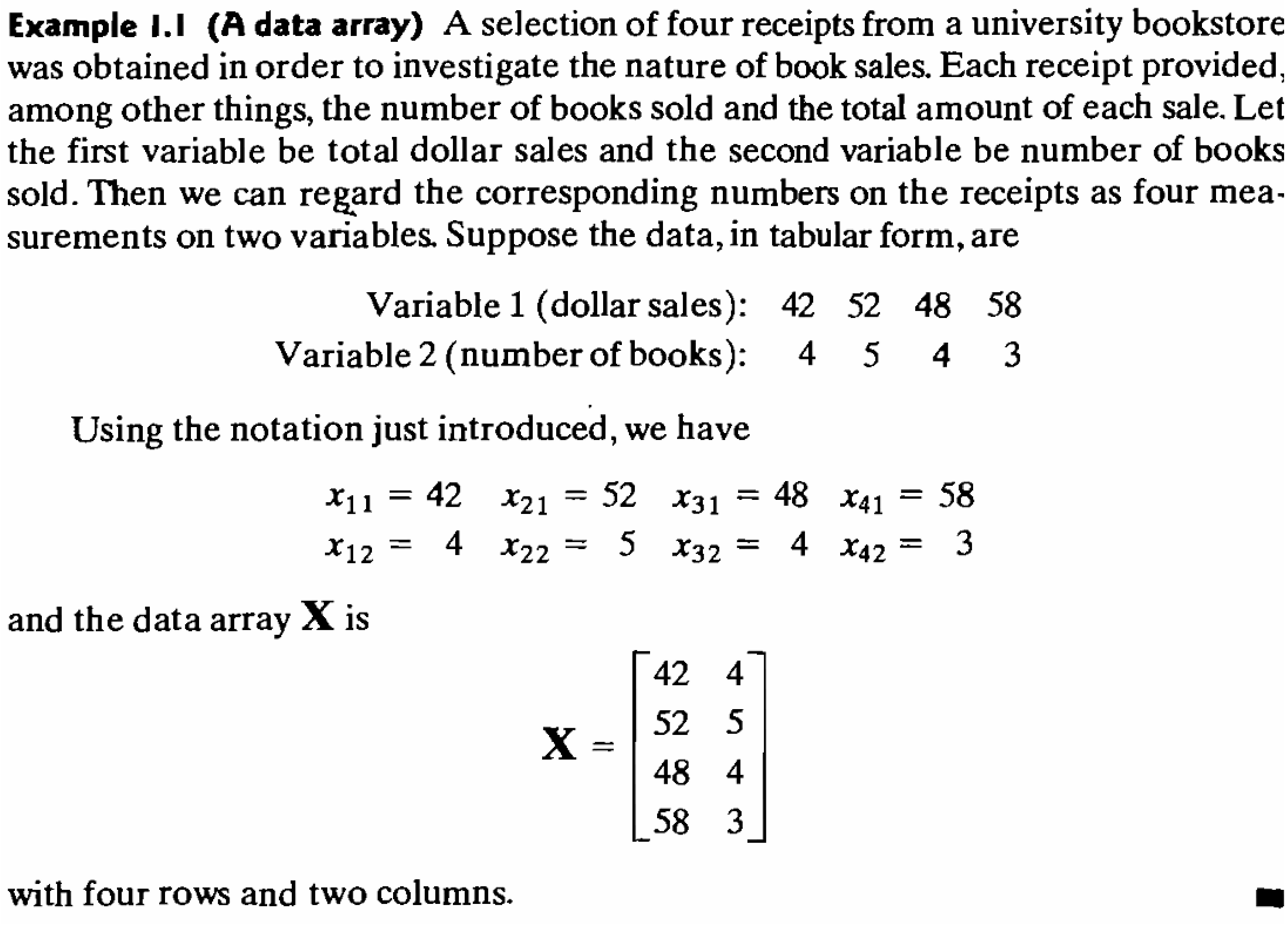
Sample correlations
$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix} \qquad r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\,\sqrt{s_{kk}}}$$

Unbiased sample covariance



$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{12} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1p} & s_{2p} & \cdots & s_{pp} \end{bmatrix} = \left\{ s_{ik} = \frac{1}{n-1}\sum_{j=1}^{n}(x_{ji}-\bar{x}_i)(x_{jk}-\bar{x}_k) \right\}$$

Example



Example 1.1 (A data array) A selection of four receipts from a university bookstore was obtained in order to investigate the nature of book sales. Each receipt provided, among other things, the number of books sold and the total amount of each sale. Let the first variable be total dollar sales and the second variable be number of books sold. Then we can regard the corresponding numbers on the receipts as four measurements on two variables. Suppose the data, in tabular form, are

Variable 1 (dollar sales):	42	52	48	58
Variable 2 (number of books):	4	5	4	3

Using the notation just introduced, we have

$$
\begin{array}{llll}
x_{11} = 42 & x_{21} = 52 & x_{31} = 48 & x_{41} = 58 \\
x_{12} = \ \ 4 & x_{22} = \ \ 5 & x_{32} = \ \ 4 & x_{42} = \ \ 3
\end{array}
$$

and the data array $\mathbf{X}$ is

$$
\mathbf{X} = \begin{bmatrix} 42 & 4 \\ 52 & 5 \\ 48 & 4 \\ 58 & 3 \end{bmatrix}
$$

with four rows and two columns. ■

The sample means are

$$\bar{x}_1 = \tfrac{1}{4}\sum_{j=1}^{4} x_{j1} = \tfrac{1}{4}(42 + 52 + 48 + 58) = 50$$

$$\bar{x}_2 = \tfrac{1}{4}\sum_{j=1}^{4} x_{j2} = \tfrac{1}{4}(4 + 5 + 4 + 3) = 4$$

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix} = \begin{bmatrix} 50 \\ 4 \end{bmatrix}$$

The sample variances and covariances are

$$\begin{aligned} s_{11} &= \tfrac{1}{4}\sum_{j=1}^{4} (x_{j1} - \bar{x}_1)^2 \\ &= \tfrac{1}{4}((42 - 50)^2 + (52 - 50)^2 + (48 - 50)^2 + (58 - 50)^2) = 34 \\ s_{22} &= \tfrac{1}{4}\sum_{j=1}^{4} (x_{j2} - \bar{x}_2)^2 \\ &= \tfrac{1}{4}((4 - 4)^2 + (5 - 4)^2 + (4 - 4)^2 + (3 - 4)^2) = .5 \\ s_{12} &= \tfrac{1}{4}\sum_{j=1}^{4} (x_{j1} - \bar{x}_1)(x_{j2} - \bar{x}_2) \\ &= \tfrac{1}{4}((42 - 50)(4 - 4) + (52 - 50)(5 - 4) \\ &\qquad + (48 - 50)(4 - 4) + (58 - 50)(3 - 4)) = -1.5 \\ s_{21} &= s_{12} \end{aligned}$$

nd

$$\mathbf{S}_n = \begin{bmatrix} 34 & -1.5 \\ -1.5 & .5 \end{bmatrix}$$

The sample correlation is

$$r_{12} = \frac{s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} = \frac{-1.5}{\sqrt{34}\sqrt{.5}} = -.36$$

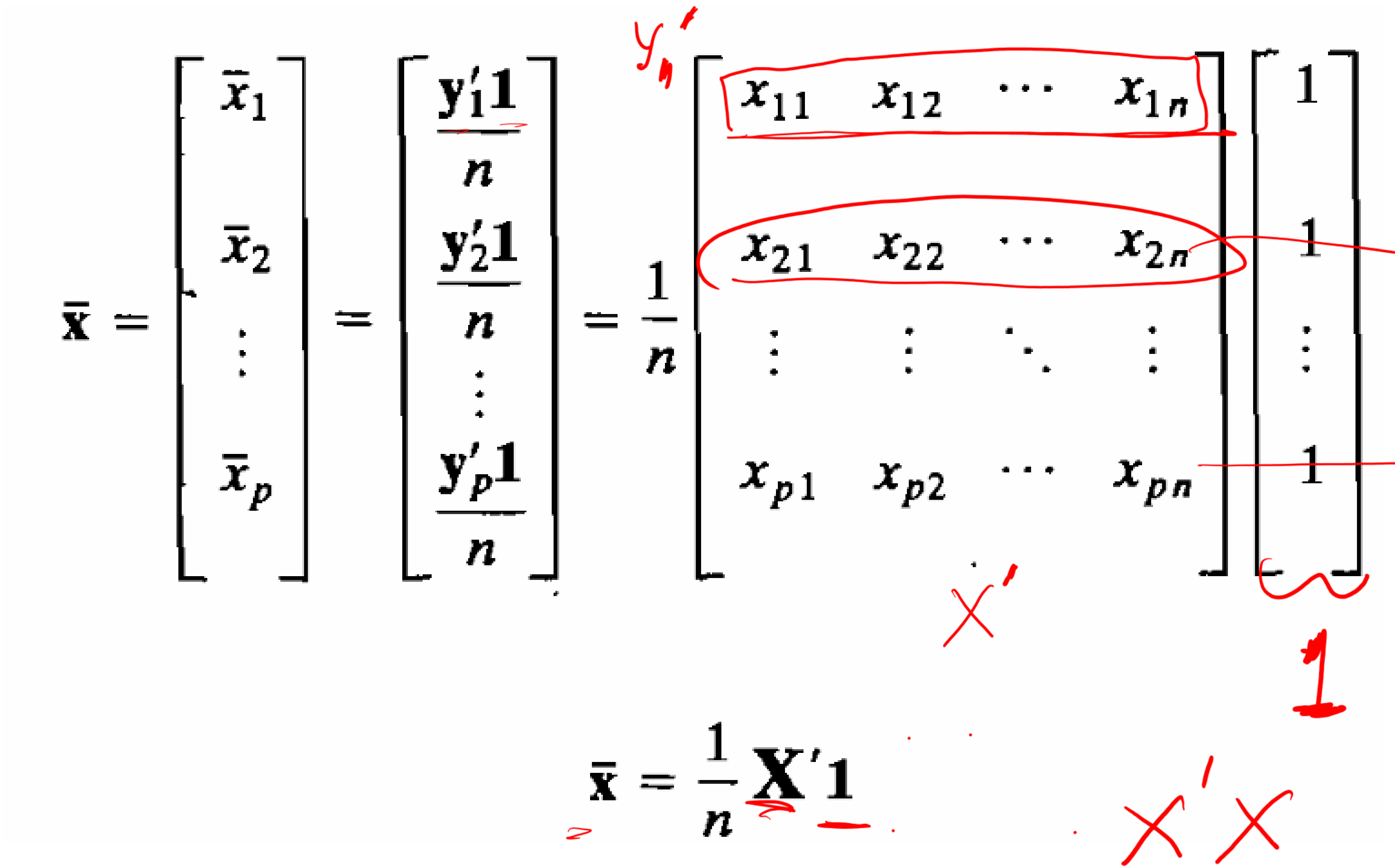
$$r_{21} = r_{12}$$

so

$$\mathbf{R} = \begin{bmatrix} 1 & -.36 \\ -.36 & 1 \end{bmatrix}$$

Descriptive statistics in matrix notation

$x_{\text{variable-id}}$, item-id



y_1'

$$\bar{\mathbf{x}} = \begin{bmatrix}\bar{x}_1\\ \bar{x}_2\\ \vdots\\ \bar{x}_p\end{bmatrix} = \begin{bmatrix}\dfrac{\mathbf{y}_1'\mathbf{1}}{n}\\ \dfrac{\mathbf{y}_2'\mathbf{1}}{n}\\ \vdots\\ \dfrac{\mathbf{y}_p'\mathbf{1}}{n}\end{bmatrix} = \frac{1}{n}\underbrace{\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n}\\ x_{21} & x_{22} & \cdots & x_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ x_{p1} & x_{p2} & \cdots & x_{pn}\end{bmatrix}}_{X'}\underbrace{\begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix}}_{\mathbf{1}}$$

$\rightarrow y_2'$

$\rightarrow y_p'$

$$\mathbf{1} \in \begin{bmatrix}1\\ \vdots\\ 1\end{bmatrix}$$

$$\bar{\mathbf{x}} = \frac{1}{n}\mathbf{X}'\mathbf{1}$$

$X_{n\times p}$

$X'X$
]

Sample covariance

$$(n-1)\underset{(p\times p)}{\mathbf{S}} = \underset{(p\times n)}{(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}')'}\underset{(n\times p)}{(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}')}$$

$$\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{n\times 1} \qquad [\bar{x}_1 \ \bar{x}_2 \ \cdots \ \bar{x}_p]$$

$$\begin{bmatrix} x_{11}-\bar{x}_1 & x_{21}-\bar{x}_1 & \cdots & x_{n1}-\bar{x}_1 \\ x_{12}-\bar{x}_2 & x_{22}-\bar{x}_2 & \cdots & x_{n2}-\bar{x}_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p}-\bar{x}_p & x_{2p}-\bar{x}_p & \cdots & x_{np}-\bar{x}_p \end{bmatrix} \begin{bmatrix} x_{11}-\bar{x}_1 & x_{12}-\bar{x}_2 & \cdots & x_{1p}-\bar{x}_p \\ x_{21}-\bar{x}_1 & x_{22}-\bar{x}_2 & \cdots & x_{2p}-\bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1}-\bar{x}_1 & x_{n2}-\bar{x}_2 & \cdots & x_{np}-\bar{x}_p \end{bmatrix}$$

$$\mathbf{1}_{n\times 1}\,\bar{x}' = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \\ \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \\ \vdots & & & \\ \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \end{bmatrix}_{n\times p}$$