Applied Multivariate Statistical Analysis

应用多元统计分析

2026 年 2 月 17 日

已 学 知 识 点

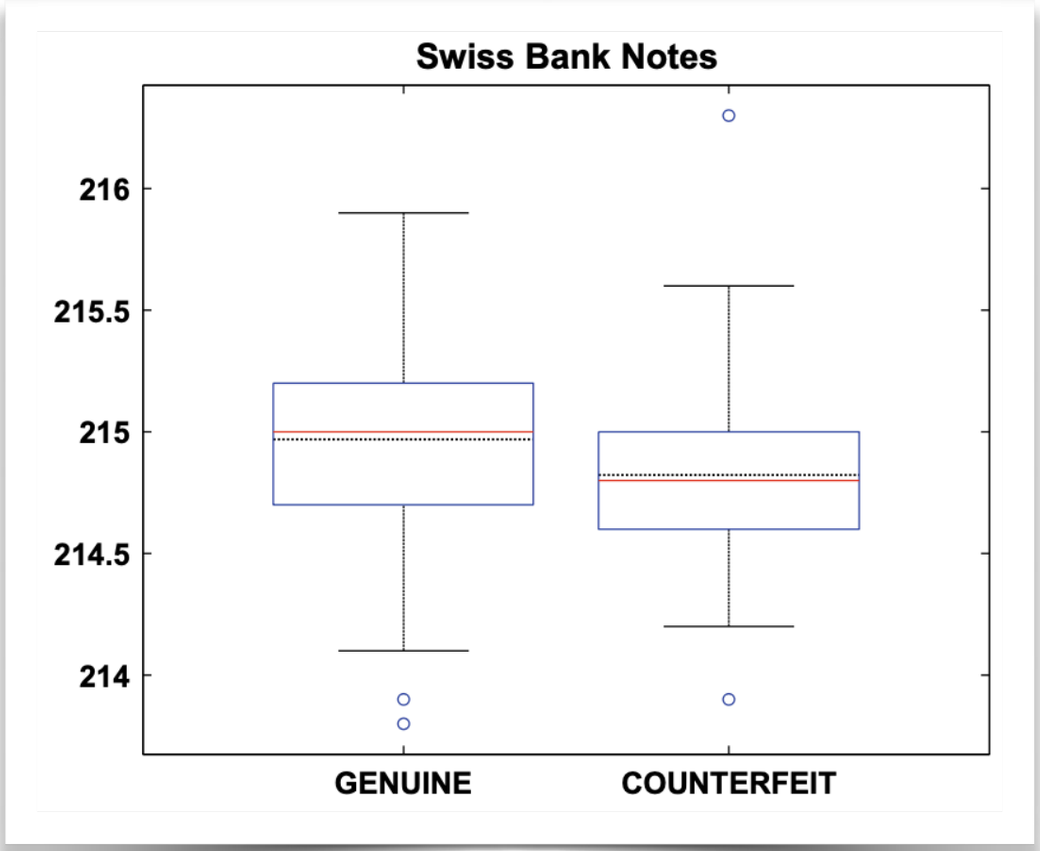
Recap

已学知识点 (Recap)

第 1 章 分类比较

1.1 箱线图

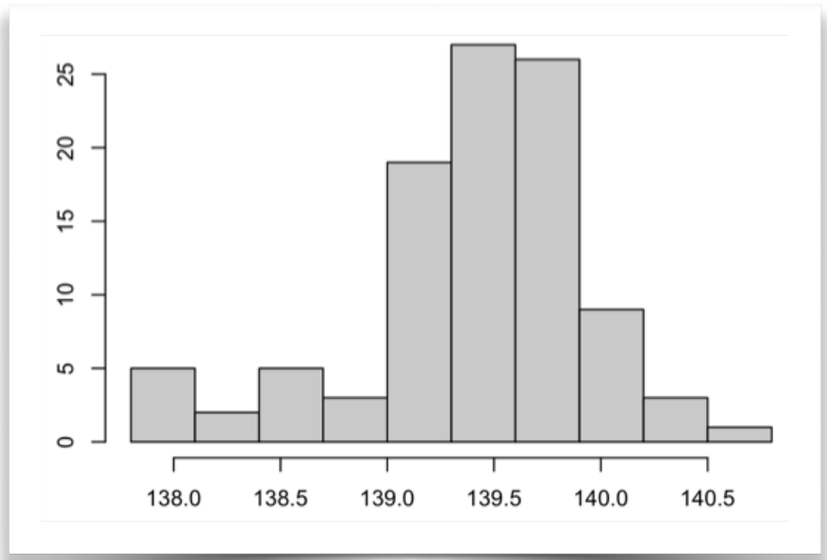
- 中位数线与均值线是对位置的度量.
- 箱的中位数线与均值线的相对位置是衡量数据集倾斜程度的指标.
- 箱与线的长度是对数据集分布范围的一种度量.
- 线的长度表示分布的尾部长度.
- 离群点 (outlying points) (可能的异常值点) 用 ○ 或 ☆ 表示，具体取决于它们是否分别位于 $F_{UL} \pm 1.5\, d_F$ 或 $F_{UL} \pm 3\, d_F$ 之外.
- 箱线图并不能表示多模态或分类.
- 关于箱的相对大小和位置的比较，实际就是对分布进行比较.



已学知识点 (Recap)

第 1 章 分类比较

1.2 直方图

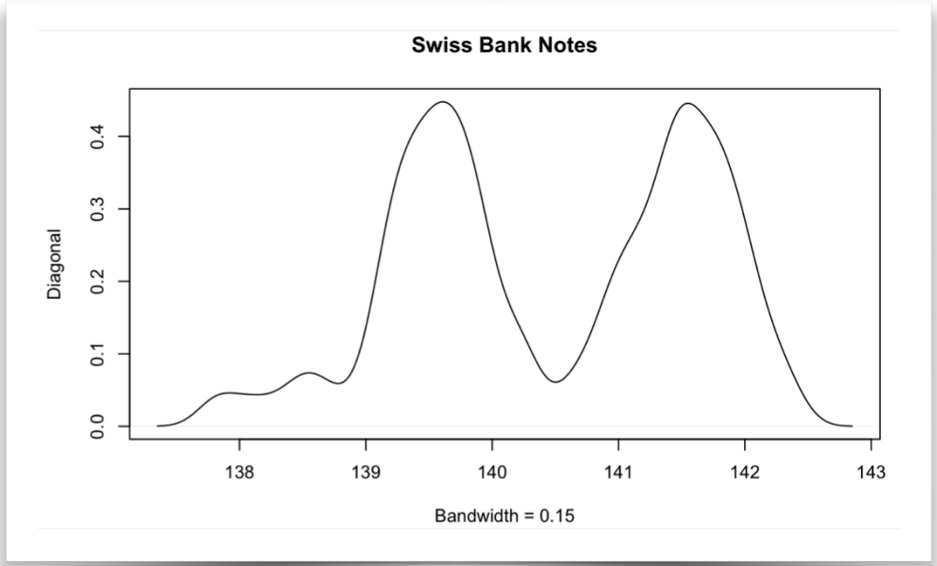


- 直方图用来检测密度的形状.
- 直方图中的强峰值对应着不同的分布特征.
- 直方图的带宽 (bin-width) h 并不要求都相等. 它也取决于分隔时的初始点 x_0.
- 初始点 x_0 的影响巨大. 改变 x_0 会看到不同形状的直方图.
- 带宽 h 的值太大会导致出现过于平坦、体现不出分布特征的直方图.
- 带宽 h 的值太小会出现极其不稳定的直方图.
- 选择最优 (optimal) 带宽 $h = \left(\frac{24\sqrt{\pi}}{n}\right)^{\frac{1}{3}}$.

已学知识点 (Recap)

第 1 章 分类比较

1.3 核密度图



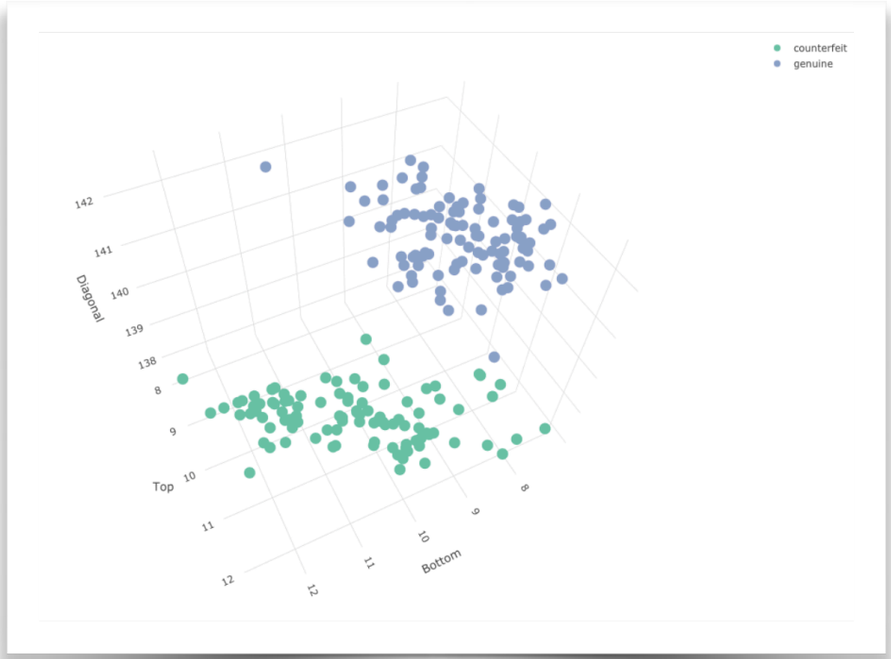
- 核密度图用核方法来估计分布的密度.
- 带宽 h 决定了估计 $\hat{f}$ 的光滑度.
- 核密度 (kernel densities) 是光滑的函数，它们可以用图形来表示分布 (最高至 3 维).
- 寻找最优带宽的一个简单 (但不总是正确) 方法是计算经验法则的带宽 $h_G = 1.06\,\hat{\sigma}\,n^{-1/5}$. 该带宽仅用于与高斯核 φ 一起.
- 核密度估计是一个很好的描述工具，可用于观察密度的形状、位置、偏度、尾部、不对称性等.

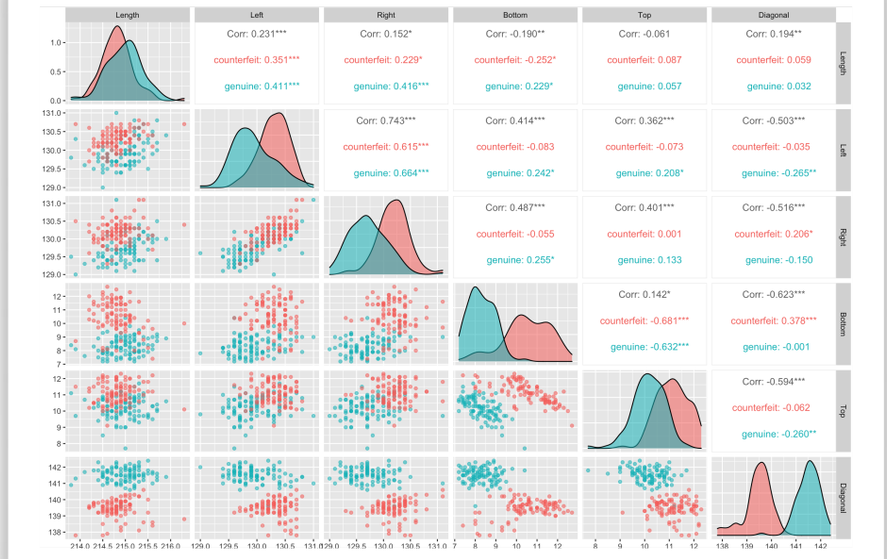
已学知识点 (Recap)

第 1 章 分类比较

1.4 散点图

- 二维和三维散点图有助于识别离群点、异常值或进行聚类.
- 散点图有助于我们判断正相关或者负相关.
- 散点图矩阵有助于我们判断其它变量值确定条件下的模型结构.

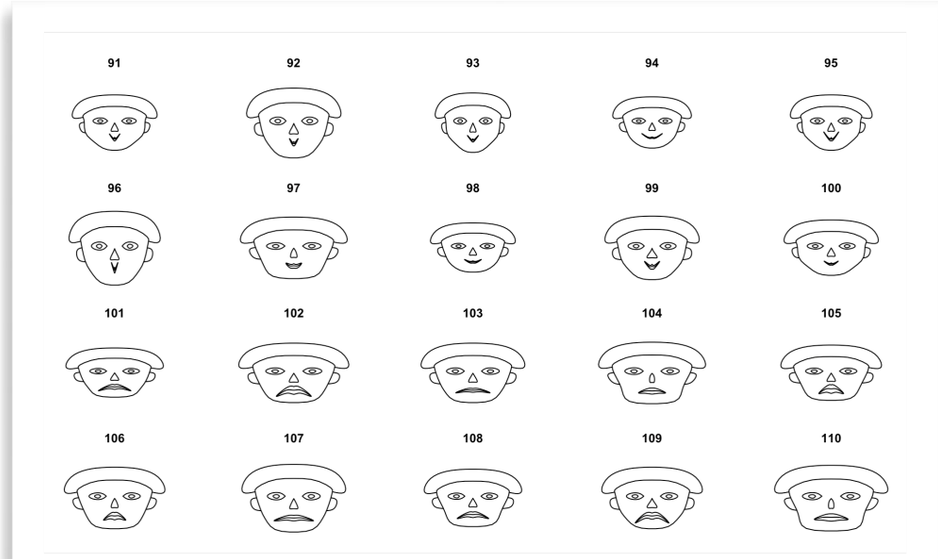




已学知识点 (Recap)

第 1 章 分类比较

1.5 脸谱图



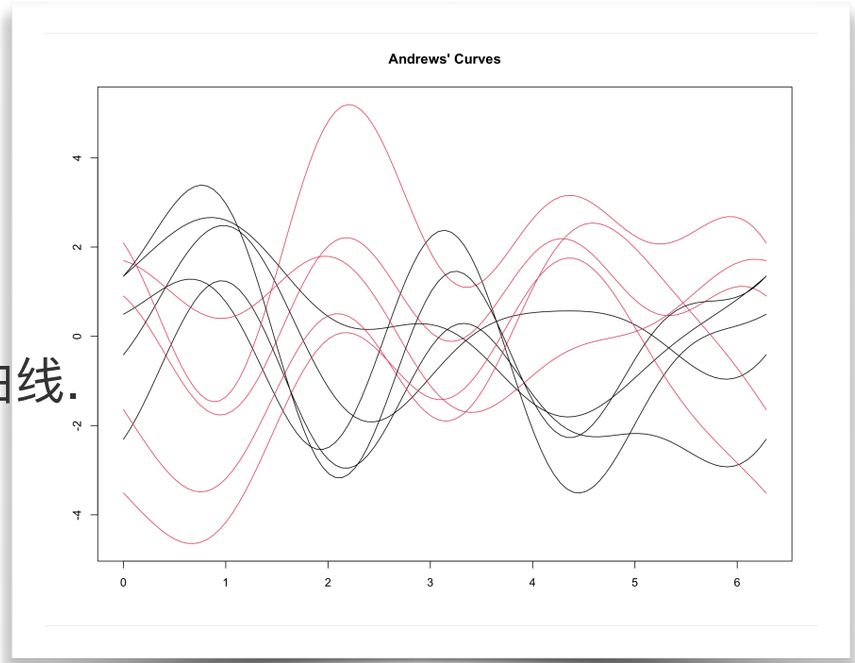
- 脸谱图可用于检测多元数据中的子群 (subgroups).
- 子群应有相似的脸谱.
- 可以通过极其不同的脸谱来判定异常值.
- 如果X的某个元素异常，则脸谱中对应于该元素的部分会出现显著的变化.

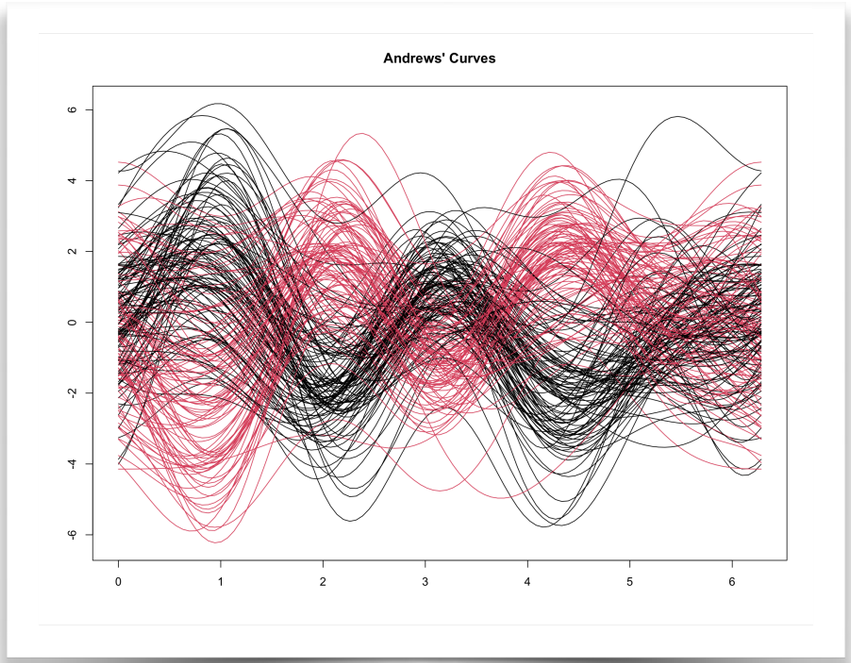
已学知识点 (Recap)

第 1 章 分类比较

1.6 Andrews 曲线

- 异常值会呈现与其它部分完全不同的一条单独的 Andrews 曲线.
- 数据集的一个子群由一组形状相似的曲线来表征.
- 变量顺序对 Andrews 曲线的解释有着重要作用.
- 变量的顺序可以通过主成分分析来优化.
- 当观测数据超过 20 次时，我们可能会得到一张糟糕的图形，一张图中有太多曲线互相覆盖.

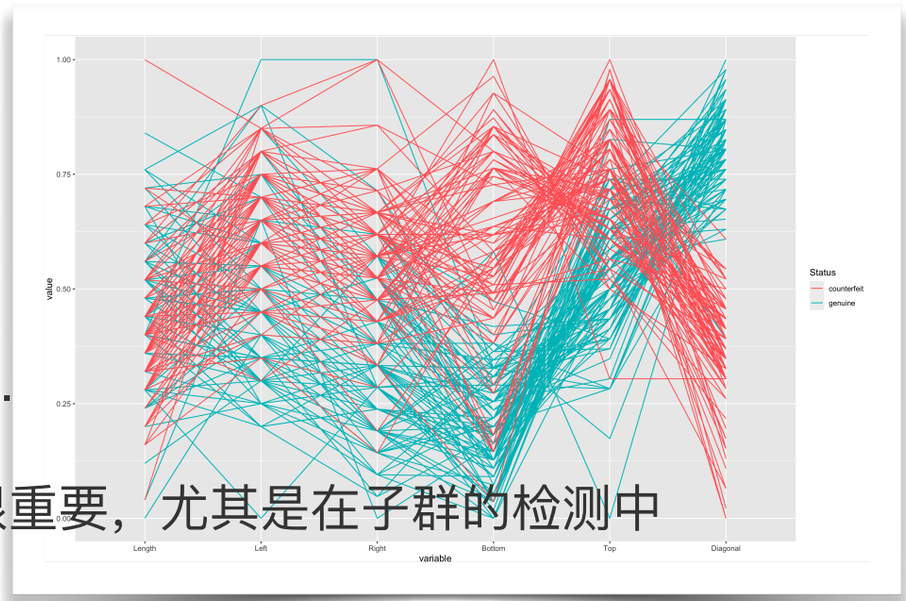




已学知识点 (Recap)

第 1 章 分类比较

1.7 平行坐标图



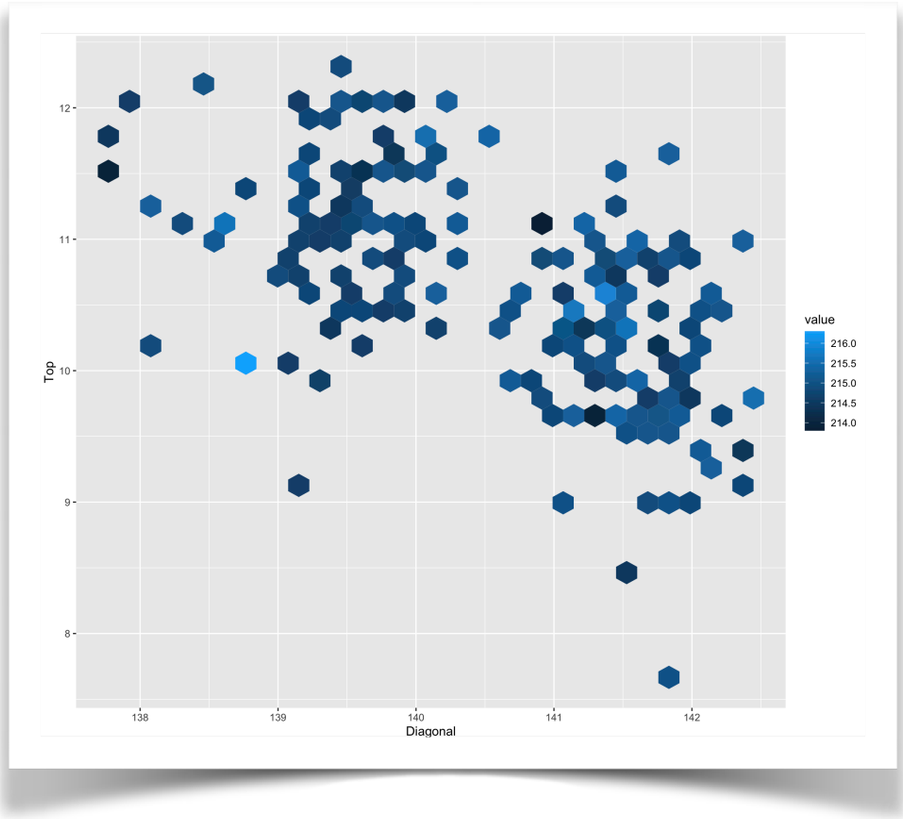
- 平行坐标图克服了维数大于 4 时直角坐标系的可视化问题.
- 异常值可以通过外围的多边形曲线表现出来.变量的顺序很重要，尤其是在子群的检测中
- 变量的顺序很重要，尤其是在检测数据的分组时.
- 分组数据可以选择不同的颜色来表示.

已学知识点 (Recap)

第 1 章 分类比较

1.8 六边形图

- 六边形图是一种以六边形为边界的二元直方图，用于大数据的可视化.
- 方差和斜度随带宽在两个相反方向变化.
- 六边形图具有“近邻对称”的性质，这是采用方形图时所缺乏的.
- 与其他镶嵌规则相比，六边形图在显示密度方面的视觉偏差较小.



已学知识点 (Recap)

第 2 章 矩阵代数初探

2.1 基本运算

- 行列式 $\left|\mathscr{A}\right|$ 的值等于 $\mathscr{A}$ 的所有特征值的乘积.
- 如果 $\left|\mathscr{A}\right| \neq 0$，则 $\mathscr{A}$ 可逆.
- 迹 $\mathrm{tr}(\mathscr{A})$ 等于 $\mathscr{A}$ 的所有特征值之和.
- 两个矩阵的迹之和等于两个矩阵之和的迹.
- $\mathrm{tr}(\mathscr{A}\mathscr{B}) = \mathrm{tr}(\mathscr{B}\mathscr{A})$.
- 矩阵 $\mathscr{A}$ 的秩 $\mathrm{rank}(\mathscr{A})$ 等于 $\mathscr{A}$ 的行 (列) 向量组的最大线性无关组中向量的个数.

已学知识点 (Recap)

第 2 章 矩阵代数初探

2.2 谱分解

$\Lambda = \mathrm{diag}\left(\lambda_1, \lambda_2, \ldots, \lambda_p\right)$ 对角矩阵

$\Gamma = \left(\gamma_1, \gamma_2, \ldots, \gamma_p\right)$ 正交矩阵

- **谱分解**：任一对称矩阵 $\mathscr{A}_{p\times p}$ 均可表示为 $\mathscr{A} = \Gamma\ \Lambda\ \Gamma^{\mathrm{T}} = \sum_{j=1}^{p} \lambda_j \gamma_j \gamma_j^{\mathrm{T}}$. 矩阵的 Jordan 分解给出了对称矩阵用其特征值和特征向量表示的一种方式.
- 对应于最大特征值的特征向量给出了数据的**主方向** (main direction).
- 矩阵的 Jordan 分解有助于我们很容易地计算对称矩阵 $\mathscr{A}$ 的幂：$\mathscr{A}^{\alpha} = \Gamma \Lambda^{\alpha} \Gamma^{\mathrm{T}}, \alpha \in \mathbb{R}$.
- **奇异值分解**：秩为 r 的任一矩阵 $\mathscr{A}_{n\times p}$ 可分解为 $\mathscr{A} = \Gamma\ \Lambda\ \Delta^{\mathrm{T}}$. 它是矩阵的 Jordan 分解在非二次形式的矩阵中的推广.

$\mathscr{A}\mathscr{A}^{\mathrm{T}}$ 的 r 个特征向量构成的矩阵

$\mathscr{A}^{\mathrm{T}}\mathscr{A}$ 对应的 r 个特征向量构成的矩阵

$\Lambda = \mathrm{diag}\left(\lambda_1^{1/2}, \lambda_2^{1/2}, \ldots, \lambda_r^{1/2}\right)$

$\lambda_1, \lambda_2, \ldots, \lambda_r$ 是矩阵 $\mathscr{A}\mathscr{A}^{\mathrm{T}}$ 与 $\mathscr{A}^{\mathrm{T}}\mathscr{A}$ 的非零特征值

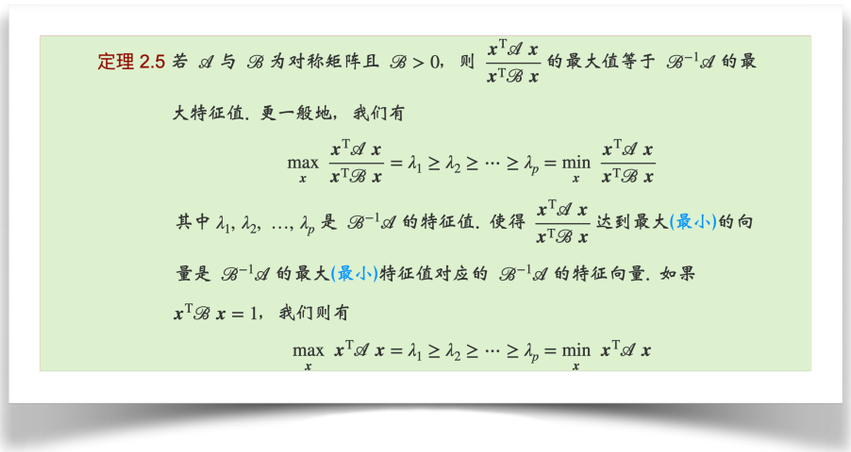
已学知识点 (Recap)

第 2 章 矩阵代数初探

2.3 二次型

- 二次型：$Q(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\mathscr{A}\,\boldsymbol{x} = \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}x_i x_j$ 可以用一个对称矩阵 $\mathscr{A}$ 描述.
- 二次型可以对角化：存在变换 $\boldsymbol{x} \mapsto \boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{y}$ 使得二次型 $\boldsymbol{x}^{\mathrm{T}}\mathscr{A}\,\boldsymbol{x} = \sum_{i=1}^{p} \lambda_i y_i^2$，其中 λ_i 是 $\mathscr{A}$ 的特征值.
- 二次型与矩阵的**正定性**：正定 (positive definite) 二次型：$\forall \boldsymbol{x} \neq \boldsymbol{0},\quad \boldsymbol{x}^{\mathrm{T}}\mathscr{A}\,\boldsymbol{x} = Q(\boldsymbol{x}) > 0$

  半正定 (positive semidefinite) 二次型：$\forall \boldsymbol{x} \neq \boldsymbol{0},\quad \boldsymbol{x}^{\mathrm{T}}\mathscr{A}\,\boldsymbol{x} = Q(\boldsymbol{x}) \geq 0$
- $\mathscr{A} > 0$ 的充分必要条件是所有 $\lambda_i > 0,\ i = 1, 2, \ldots, p.$
- 给定某些约束条件的二次型的最大值和最小值可以用特征值来表示.



定理 2.5 若 $\mathscr{A}$ 与 $\mathscr{B}$ 为对称矩阵且 $\mathscr{B} > 0$，则 $\frac{\boldsymbol{x}^{\mathrm{T}}\mathscr{A}\,\boldsymbol{x}}{\boldsymbol{x}^{\mathrm{T}}\mathscr{B}\,\boldsymbol{x}}$ 的最大值等于 $\mathscr{B}^{-1}\mathscr{A}$ 的最大特征值. 更一般地，我们有

$$\max_{\boldsymbol{x}} \frac{\boldsymbol{x}^{\mathrm{T}}\mathscr{A}\,\boldsymbol{x}}{\boldsymbol{x}^{\mathrm{T}}\mathscr{B}\,\boldsymbol{x}} = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p = \min_{\boldsymbol{x}} \frac{\boldsymbol{x}^{\mathrm{T}}\mathscr{A}\,\boldsymbol{x}}{\boldsymbol{x}^{\mathrm{T}}\mathscr{B}\,\boldsymbol{x}}$$

其中 $\lambda_1, \lambda_2, \ldots, \lambda_p$ 是 $\mathscr{B}^{-1}\mathscr{A}$ 的特征值. 使得 $\frac{\boldsymbol{x}^{\mathrm{T}}\mathscr{A}\,\boldsymbol{x}}{\boldsymbol{x}^{\mathrm{T}}\mathscr{B}\,\boldsymbol{x}}$ 达到最大(最小)的向量是 $\mathscr{B}^{-1}\mathscr{A}$ 的最大(最小)特征值对应的 $\mathscr{B}^{-1}\mathscr{A}$ 的特征向量. 如果 $\boldsymbol{x}^{\mathrm{T}}\mathscr{B}\,\boldsymbol{x} = 1$，我们则有

$$\max_{\boldsymbol{x}} \boldsymbol{x}^{\mathrm{T}}\mathscr{A}\,\boldsymbol{x} = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p = \min_{\boldsymbol{x}} \boldsymbol{x}^{\mathrm{T}}\mathscr{A}\,\boldsymbol{x}$$

已学知识点 (Recap)

第 2 章 矩阵代数初探

2.4 导数

- f 的梯度：$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} \triangleq \begin{pmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} \\ \frac{\partial f(\boldsymbol{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\boldsymbol{x})}{\partial x_p} \end{pmatrix}, \quad \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}^{\mathrm{T}}} \triangleq \begin{pmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1}, & \frac{\partial f(\boldsymbol{x})}{\partial x_2}, & \cdots, & \frac{\partial f(\boldsymbol{x})}{\partial x_p} \end{pmatrix}$$

- f 的 Hessian 矩阵：$$\frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^{\mathrm{T}}} \triangleq \begin{pmatrix} \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_1} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_p} \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_p \partial x_1} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_p \partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_p \partial x_p} \end{pmatrix}$$

已学知识点 (Recap)

第 2 章 矩阵代数初探

2.5 分块矩阵

$$\mathscr{A}_{n\times p}=\begin{pmatrix}\mathscr{A}_{11} & \mathscr{A}_{12}\\ \mathscr{A}_{21} & \mathscr{A}_{22}\end{pmatrix} \overset{n=p}{\Longrightarrow} \mathscr{A}_{p\times p}=\begin{pmatrix}\mathscr{A}_{11} & \mathscr{A}_{12}\\ \mathscr{A}_{21} & \mathscr{A}_{22}\end{pmatrix}$$

方阵

- 逆矩阵：如果 $\mathscr{A}$ 非奇异 $\left(\mathscr{A}\mathscr{A}^{-1}=\mathscr{I}_p\right) \Longrightarrow \mathscr{A}^{-1}=\begin{pmatrix}\mathscr{A}^{11} & \mathscr{A}^{12}\\ \mathscr{A}^{21} & \mathscr{A}^{22}\end{pmatrix}$

$$|\mathscr{A}|=\begin{cases}|\mathscr{A}_{11}|\,|\mathscr{A}_{22}-\mathscr{A}_{21}\mathscr{A}_{11}^{-1}\mathscr{A}_{12}|=|\mathscr{A}_{11}|\,|\mathscr{A}_{22\cdot 1}|, & \mathscr{A}_{11}\text{非奇异}\\ |\mathscr{A}_{22}|\,|\mathscr{A}_{11}-\mathscr{A}_{12}\mathscr{A}_{22}^{-1}\mathscr{A}_{21}|=|\mathscr{A}_{22}|\,|\mathscr{A}_{11\cdot 2}|, & \mathscr{A}_{22}\text{非奇异}\end{cases}$$

$$\begin{cases}\mathscr{A}^{11}=\left(\mathscr{A}_{11}-\mathscr{A}_{12}\mathscr{A}_{22}^{-1}\mathscr{A}_{21}\right)^{-1}\overset{\text{def}}{=}\left(\mathscr{A}_{11\cdot 2}\right)^{-1}\\ \mathscr{A}^{12}=-\left(\mathscr{A}_{11\cdot 2}\right)^{-1}\mathscr{A}_{12}\mathscr{A}_{22}^{-1}\\ \mathscr{A}^{21}=-\mathscr{A}_{22}^{-1}\mathscr{A}_{21}\left(\mathscr{A}_{11\cdot 2}\right)^{-1}\\ \mathscr{A}^{22}=\mathscr{A}_{22}^{-1}+\mathscr{A}_{22}^{-1}\mathscr{A}_{21}\left(\mathscr{A}_{11\cdot 2}\right)^{-1}\mathscr{A}_{12}\mathscr{A}_{22}^{-1}\end{cases}$$

$$\begin{cases}\mathscr{A}_{11\cdot 2}\overset{\text{def}}{=}\mathscr{A}_{11}-\mathscr{A}_{12}\mathscr{A}_{22}^{-1}\mathscr{A}_{21}\\ \mathscr{A}_{22\cdot 1}\overset{\text{def}}{=}\mathscr{A}_{22}-\mathscr{A}_{21}\mathscr{A}_{11}^{-1}\mathscr{A}_{12}\end{cases}$$

- 设 $\mathscr{A}_{n\times p}$ 为 $n\times p$ 矩阵，$\mathscr{B}_{p\times n}$ 为 $p\times n$ 矩阵，则 $\mathscr{A}\mathscr{B}$ 与 $\mathscr{B}\mathscr{A}$ 有相同的非零特征值，且特征值的重数亦相同. 若 $\boldsymbol{x}$ 是矩阵 $\mathscr{A}\mathscr{B}$ 对应某个非零特征值 $\lambda\neq 0$ 的一个特征向量，则 $\boldsymbol{y}=\mathscr{B}\boldsymbol{x}$ 是矩阵 $\mathscr{B}\mathscr{A}$ 对应该特征值的一个特征向量.

几何直观 (geometrical aspects)

- 距离 (distance)
  - 定义: $\mathbb{R}^p$ 中的两点 $\boldsymbol{x}, \boldsymbol{y}$ 的**距离** $d(\boldsymbol{x}, \boldsymbol{y})$ 定义为函数 $d: \mathbb{R}^{2p} \to \mathbb{R}_+$，它满足

$$\begin{cases} d(\boldsymbol{x}, \boldsymbol{y}) > 0, & \forall \boldsymbol{x} \neq \boldsymbol{y} \\ d(\boldsymbol{x}, \boldsymbol{y}) = 0, & \text{当且仅当 } \boldsymbol{x} = \boldsymbol{y} \\ d(\boldsymbol{x}, \boldsymbol{y}) \leq d(\boldsymbol{x}, \boldsymbol{z}) + d(\boldsymbol{z}, \boldsymbol{y}), & \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \end{cases}$$

  - 欧氏距离 (Euclidean distance):

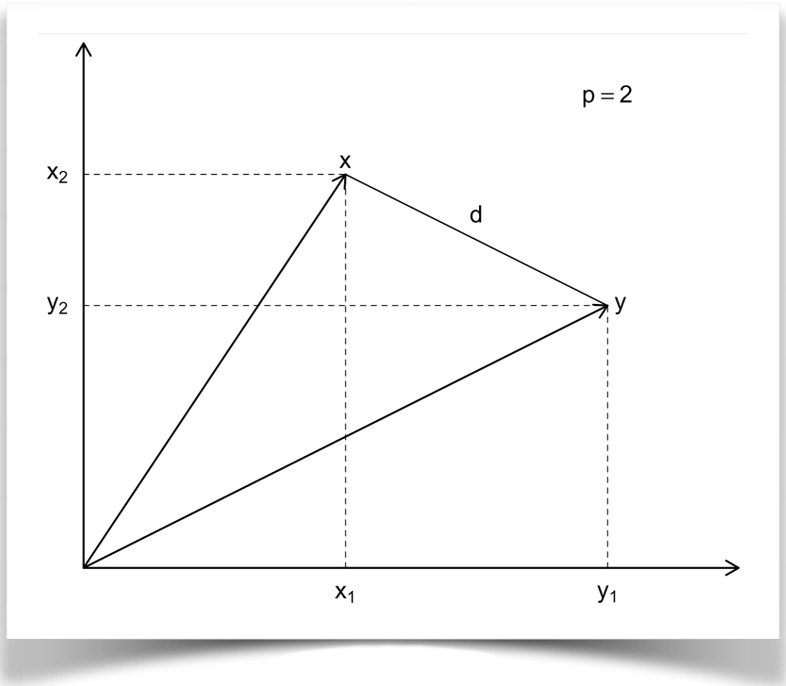
$$d^2(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} - \boldsymbol{y})^{\mathrm{T}} \mathscr{A} (\boldsymbol{x} - \boldsymbol{y})$$

正定矩阵 ($\mathscr{A} > 0$)

度量 (metric)

$$\mathscr{A} = \mathscr{I}_p$$

$$d^2(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{p} (x_i - y_i)^2$$

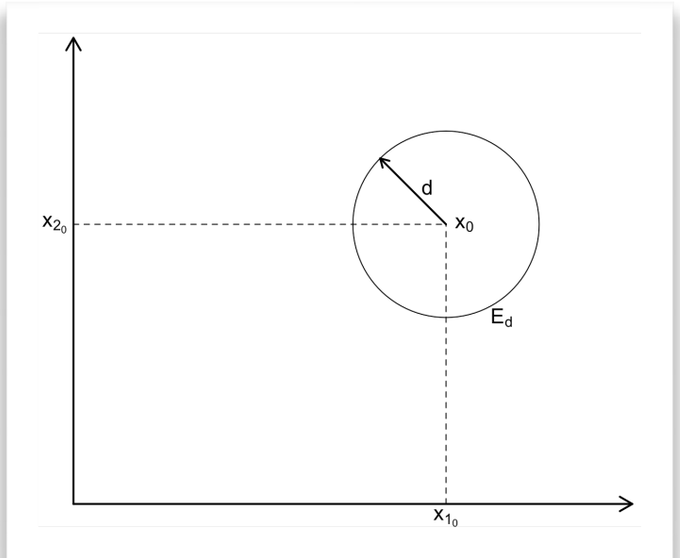


几何直观 (geometrical aspects)

- 距离 (distance)
  - 欧氏 $\mathscr{I}_p$ 距离，到点 $\boldsymbol{x}_0$ 的**等距曲线** (iso-distance curve)

$$E_d = \left\{ \boldsymbol{x} \in \mathbb{R}^p \,\middle|\, (\boldsymbol{x}-\boldsymbol{x}_0)^{\mathrm{T}} (\boldsymbol{x}-\boldsymbol{x}_0) = d^2 \right\}$$

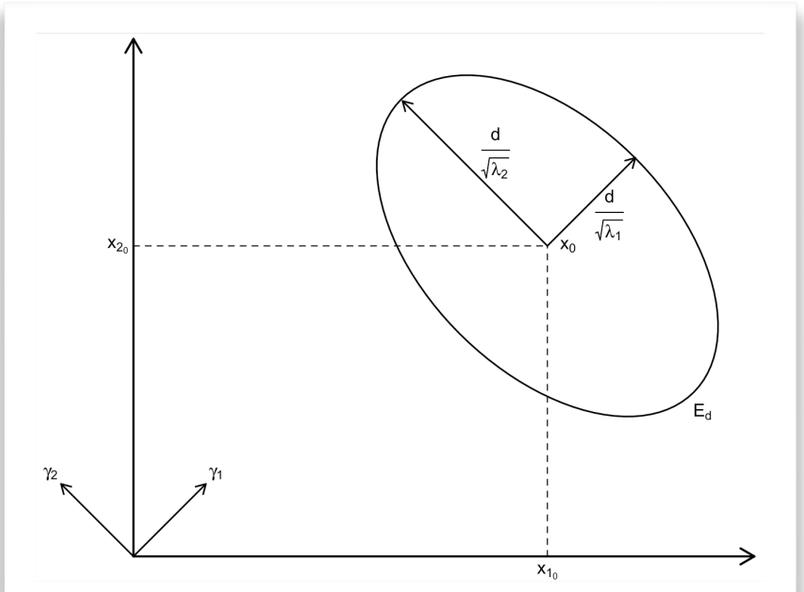
中心点为 $\boldsymbol{x}_0$ 的球面，其中 d 为常数.



  - 一般的等距曲线

$$E_d = \left\{ \boldsymbol{x} \in \mathbb{R}^p \,\middle|\, (\boldsymbol{x}-\boldsymbol{x}_0)^{\mathrm{T}} \mathscr{A} (\boldsymbol{x}-\boldsymbol{x}_0) = d^2 \right\}$$

中心点为 $\boldsymbol{x}_0$ 的椭球面，其中 $\mathscr{A} > 0$ 为正定矩阵，d 为常数.



几何直观 (geometrical aspects)

- **定理 2.7** 设 $\gamma_1, \gamma_2, \ldots, \gamma_p$ 是矩阵 $\mathscr{A}$ 的特征值 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ 对应的正交特征向量.

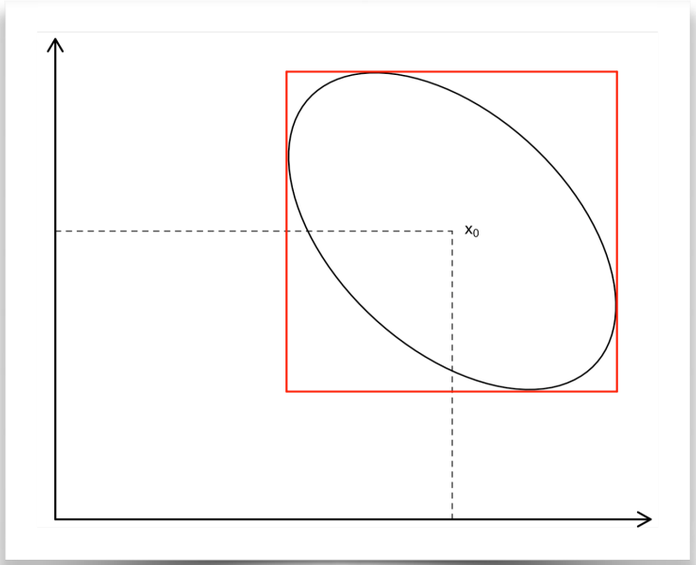
  (i) 椭球面 E_d 的主轴位于 γ_i , $i = 1, 2, \ldots, p$ 的方向.

  (ii) 半轴的长度为 $\sqrt{\dfrac{d^2}{\lambda_i}}$, $i = 1, 2, \ldots, p$.

  (iii) 包围椭球体 E_d 的外接矩形由以下不等式确定:

$$x_{0i} - \sqrt{d^2 a^{ii}} \leq x_i \leq x_{0i} + \sqrt{d^2 a^{ii}} \ , \quad i = 1, 2, \ldots, p$$

  其中 a^{ii} 是 $\mathscr{A}^{-1}$ 的 (i,i) 元素. 包围椭球体 E_d 的外接矩形我们指的是边与坐标轴平行的矩形.

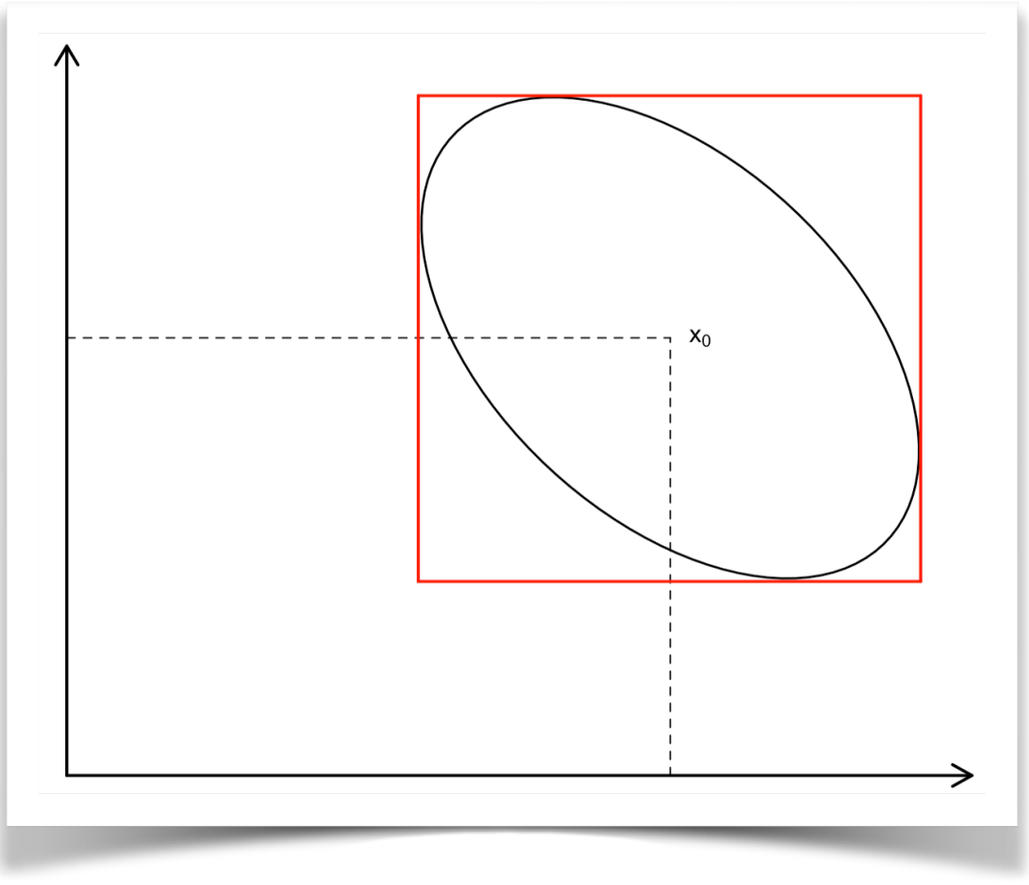


几何直观 (geometrical aspects)

- 不难确定椭球体与其平行于坐标轴的外接矩形的切点.
- 我们来确定第 j 个坐标轴方向 (正方向) 上的切点坐标.
- 为简单起见，我们假设椭球以原点 $\left(\boldsymbol{x}_0 = \mathbf{0}\right)$ 为中心. 否则，只需将矩形平移 $\boldsymbol{x}_0$ 即可.
- 切点的坐标由以下问题的解给出:

$$\boldsymbol{x} = \arg \max_{\boldsymbol{x}^{\mathrm{T}}\mathscr{A}\boldsymbol{x}=d^2} \boldsymbol{e}_j^{\mathrm{T}}\boldsymbol{x}$$

单位矩阵 $\mathscr{I}_p$ 的第 j 列



几何直观 (geometrical aspects)

- 可通过拉格朗日乘数法来求解

$$L = \boldsymbol{e}_j^{\mathrm{T}}\boldsymbol{x} - \lambda\left(\boldsymbol{x}^{\mathrm{T}}\mathscr{A}\boldsymbol{x} - d^2\right)$$

$$\begin{cases} \dfrac{\partial L}{\partial \boldsymbol{x}} = \boldsymbol{e}_j - 2\lambda\mathscr{A}\boldsymbol{x} = \boldsymbol{0} \\ \dfrac{\partial L}{\partial \lambda} = \boldsymbol{x}^{\mathrm{T}}\mathscr{A}\boldsymbol{x} - d^2 = 0 \end{cases}$$

$$\boldsymbol{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix}, \quad \mathscr{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{pmatrix} = \mathscr{A}^{\mathrm{T}}, \quad \boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$$

$$\Longrightarrow \quad \frac{\partial \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{x}^{\mathrm{T}}\boldsymbol{a}}{\partial \boldsymbol{x}} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \boldsymbol{a}, \qquad \frac{\partial \boldsymbol{x}^{\mathrm{T}}\mathscr{A}\,\boldsymbol{x}}{\partial \boldsymbol{x}} = 2\mathscr{A}\boldsymbol{x}$$

$$\Longrightarrow \quad \boldsymbol{x} = \frac{1}{2\lambda}\mathscr{A}^{-1}\boldsymbol{e}_j \quad \Longrightarrow \quad x_i = \frac{1}{2\lambda}a^{ij}, \; i = 1, 2, \ldots, p$$

a^{ij}：$\mathscr{A}^{-1}$ 的 (i, j) 元素

几何直观 (geometrical aspects)

$$\begin{cases} \dfrac{\partial L}{\partial \boldsymbol{x}} = \boldsymbol{e}_j - 2\lambda \mathscr{A}\boldsymbol{x} = \boldsymbol{0} \\ \dfrac{\partial L}{\partial \lambda} = \boldsymbol{x}^{\mathrm{T}}\mathscr{A}\boldsymbol{x} - d^2 = 0 \end{cases}$$

左乘 $\boldsymbol{x}^{\mathrm{T}}$

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{e}_j - 2\lambda \boldsymbol{x}^{\mathrm{T}}\mathscr{A}\boldsymbol{x} = \boldsymbol{0}$$

$$\Longrightarrow \quad \boldsymbol{x}^{\mathrm{T}}\boldsymbol{e}_j = 2\lambda d^2 \quad \Longrightarrow \quad x_j = 2\lambda d^2$$

$$x_i = \frac{1}{2\lambda} a^{ij}, \ i = 1, 2, \ \ldots, p$$

(a^{ij}：$\mathscr{A}^{-1}$ 的 (i, j) 元素)

$i = j$

$$\frac{1}{2\lambda} a^{jj} = 2\lambda d^2 \quad \Longrightarrow \quad 2\lambda = \sqrt{\frac{a^{jj}}{d^2}}$$

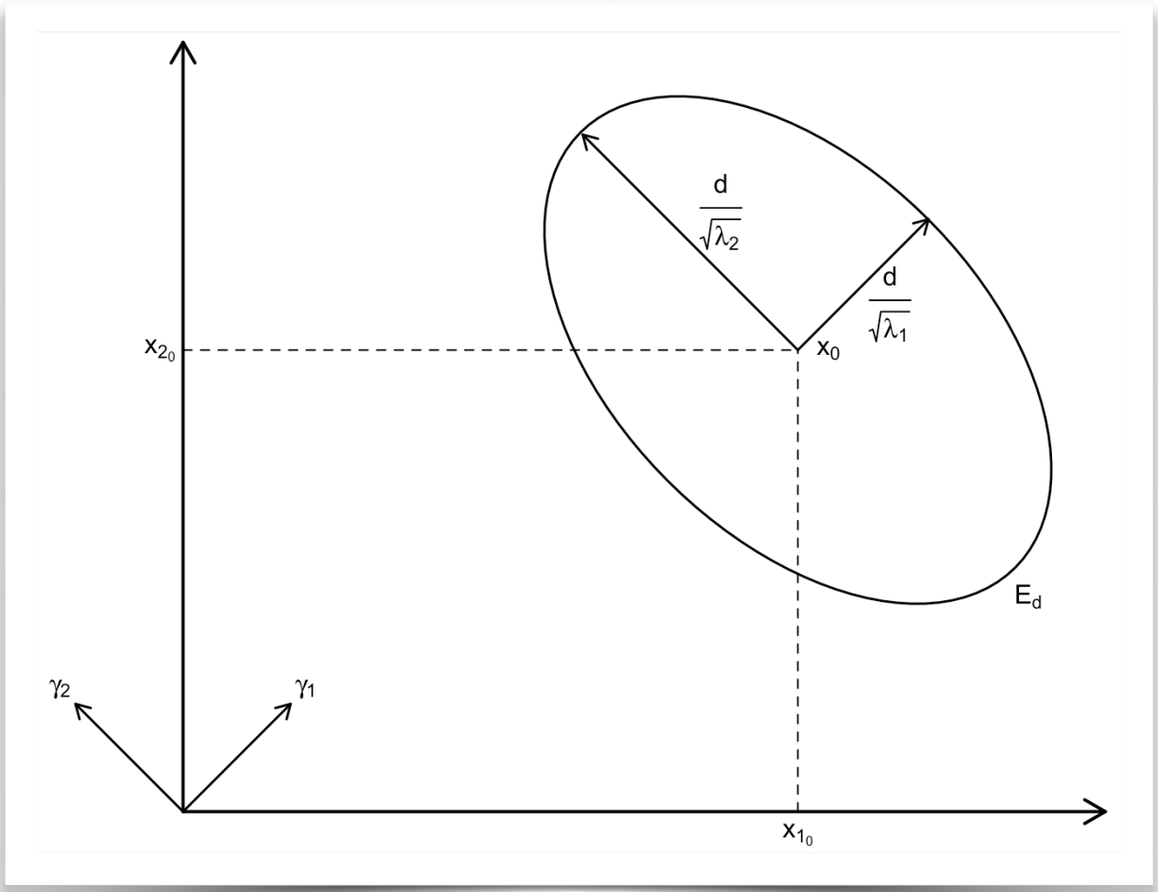
- 椭球面与包围它的矩形在第 j 个轴的正方向上的切点坐标为

$$x_i = \sqrt{\frac{d^2}{a^{jj}}}\, a^{ij}, \quad i = 1, 2, \ \ldots, p$$

$i = j$ → 定理 2.7 (*iii*)

几何直观 (geometrical aspects)

- 定理 2.7 的作用
  - 首先，它为绘制二维椭圆提供了一个有用的工具.
  - 可以证明，多元正态总体均值向量 $\boldsymbol{\mu}$ 的置信域可由一个特定的椭球体给出，椭球的参数取决于样本的特征.
  - 还将证明，多元正态分布密度的等值面亦由椭球体给出，椭球体的参数取决于多元正态分布的均值向量和协方差矩阵.



几何直观 (geometrical aspects)

- 向量的范数 (norm)
  - 给定 $\mathbb{R}^p$ 中的一个向量 $\boldsymbol{x}$，定义 $\boldsymbol{x}$ (关于度量 $\mathscr{I}_p$) 的范数或长度 (length) 为

$$\|\boldsymbol{x}\| = d\left(\mathbf{0}_p, \boldsymbol{x}\right) = \sqrt{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}}$$

  - 当 $\|\boldsymbol{x}\| = 1$ 时，称 $\boldsymbol{x}$ 为单位向量 (unit vector).
  - 更一般地，$\boldsymbol{x}$ (关于度量 $\mathscr{A}$) 的范数定义为

$$\|\boldsymbol{x}\|_{\mathscr{A}} = \sqrt{\boldsymbol{x}^{\mathrm{T}}\mathscr{A}\boldsymbol{x}}$$

几何直观 (geometrical aspects)

- 两个向量的夹角
  - 考虑两个向量 $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$，$\boldsymbol{x}$ 与 $\boldsymbol{y}$ 的夹角 θ 可由其余弦如下定义：

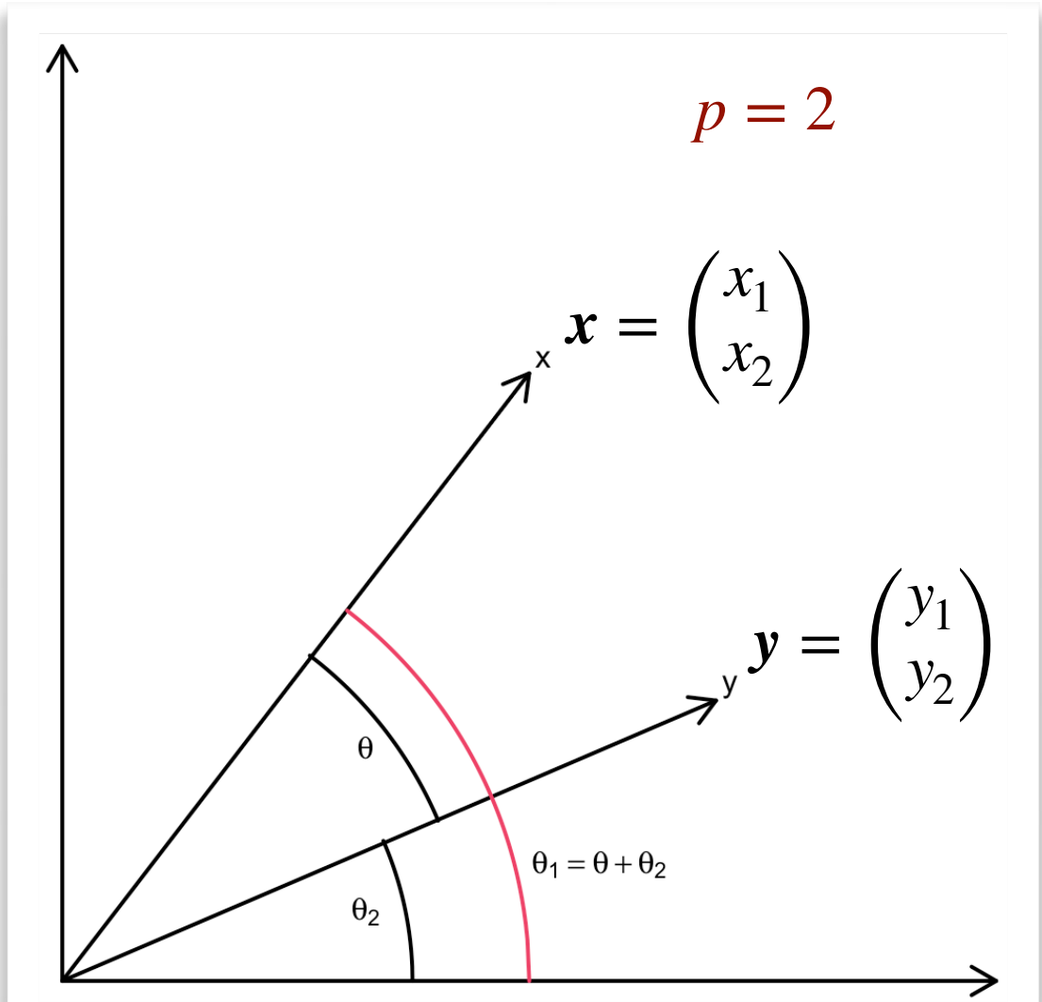
$$\cos\theta = \frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}}{\|\boldsymbol{x}\|\ \|\boldsymbol{y}\|}$$

$$\begin{cases} \|\boldsymbol{x}\|\cos\theta_1 = x_1 \\ \|\boldsymbol{x}\|\sin\theta_1 = x_2 \end{cases} \qquad \begin{cases} \|\boldsymbol{y}\|\cos\theta_2 = y_1 \\ \|\boldsymbol{y}\|\sin\theta_2 = y_2 \end{cases}$$

$$\Longrightarrow \cos\theta = \cos(\theta_1 - \theta_2)$$

$$= \cos\theta_1\cos\theta_2 + \sin\theta_1\sin\theta_2$$

$$= \frac{x_1}{\|\boldsymbol{x}\|}\frac{y_1}{\|\boldsymbol{y}\|} + \frac{x_2}{\|\boldsymbol{x}\|}\frac{y_2}{\|\boldsymbol{y}\|} = \frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}}{\|\boldsymbol{x}\|\ \|\boldsymbol{y}\|}$$



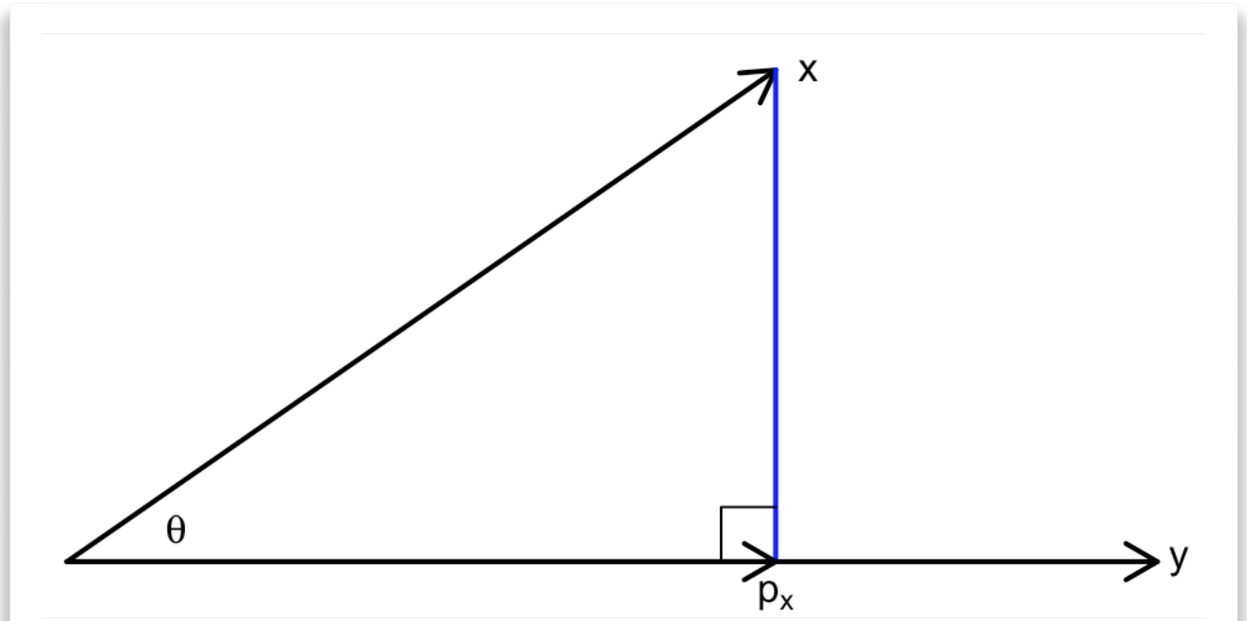
几何直观 (geometrical aspects)

- 注 2.1 如果 $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}=0$, 则夹角 $\theta=\dfrac{\pi}{2}$.

$$\cos\theta=\frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}}{\|\boldsymbol{x}\|\ \|\boldsymbol{y}\|}$$

$$\|\boldsymbol{p}_x\| = \|\boldsymbol{x}\|\cdot|\cos\theta| = \frac{\left|\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}\right|}{\|\boldsymbol{y}\|}$$

$\boldsymbol{x}$ 在 $\boldsymbol{y}$ 上的投影: $\boldsymbol{p}_x=\dfrac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}}{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}}\boldsymbol{y}$



- 关于一般度量 $\mathscr{A}$ 也可以定义向量的夹角

$$\cos\theta=\frac{\boldsymbol{x}^{\mathrm{T}}\mathscr{A}\boldsymbol{y}}{\|\boldsymbol{x}\|_{\mathscr{A}}\cdot\|\boldsymbol{y}\|_{\mathscr{A}}}$$

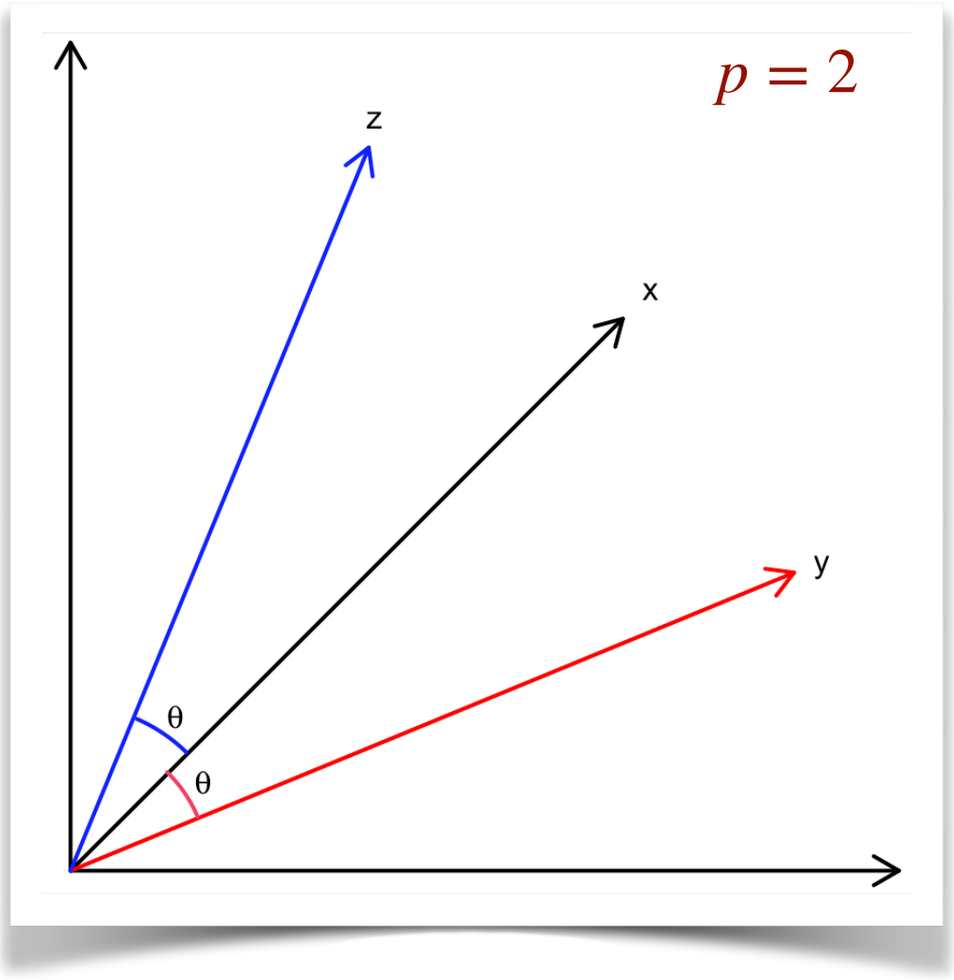
- 如果 $\cos\theta=0$, 则称 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 关于度量 $\mathscr{A}$ 正交.

几何直观 (geometrical aspects)

- 旋转 (rotation)

正交矩阵：$\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathrm{T}} = \boldsymbol{I}$

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \boldsymbol{\Gamma} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$



- 关于原点顺时针旋转一个角度 θ

$$\boldsymbol{y} = \boldsymbol{\Gamma}\,\boldsymbol{x}$$

- 关于原点逆时针旋转一个角度 θ

$$\boldsymbol{z} = \boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{x}$$

- 一般地，向量 $\boldsymbol{x}$ 左乘一个正交矩阵 $\boldsymbol{\Gamma}$，在几何上对应于坐标轴的旋转，从而第一个新轴由 $\boldsymbol{\Gamma}$ 的第一行确定.

几何直观 (geometrical aspects)

- 矩阵的列空间 (Column Space) 与零空间 (Null Space)

  - 对矩阵 $\mathcal{X}_{n\times p}$，定义 $\mathcal{X}$ 的列空间，或 $\mathcal{X}$ 的列向量生成的空间为

$$C(\mathcal{X}) \triangleq \left\{ \boldsymbol{x} \in \mathbb{R}^n \,\middle|\, \text{存在向量} \boldsymbol{a} \in \mathbb{R}^p \text{使得} \mathcal{X}\boldsymbol{a} = \boldsymbol{x} \right\}$$

$$\Longrightarrow \quad C(\mathcal{X}) \subseteq \mathbb{R}^n, \quad \dim\left\{ C(\mathcal{X}) \right\} = \text{rank}(\mathcal{X}) = r \leq \min\{n, p\}$$

  - $\mathcal{X}$ 的零空间定义为

$$N(\mathcal{X}) \triangleq \left\{ \boldsymbol{y} \in \mathbb{R}^p \,\middle|\, \mathcal{X}\boldsymbol{y} = \boldsymbol{0} \right\}$$

$$\Longrightarrow \quad N(\mathcal{X}) \subseteq \mathbb{R}^p, \quad \dim\left\{ N(\mathcal{X}) \right\} = p - r$$

几何直观 (geometrical aspects)

- 注 2.2 $N\left(\mathcal{X}^{\mathrm{T}}\right)$ 是 $\mathbb{R}^n$ 中 $C\left(\mathcal{X}\right)$ 的正交补 (orthogonal complement)，即，给定向量 $\boldsymbol{b} \in \mathbb{R}^n$，

  则 $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{b} = 0$ 对所有 $\boldsymbol{x} \in C\left(\mathcal{X}\right)$ 都成立的充分必要条件是 $\boldsymbol{b} \in N\left(\mathcal{X}^{\mathrm{T}}\right)$.

- 设 $\mathcal{X} = \begin{pmatrix} 2 & 3 & 5 \\ 4 & 6 & 7 \\ 6 & 8 & 6 \\ 8 & 2 & 4 \end{pmatrix}$.

$$\Longrightarrow \quad \operatorname{rank}\left(\mathcal{X}\right) = 3. \quad \Longrightarrow \quad \dim\left\{C\left(\mathcal{X}\right)\right\} = 3.$$

$$\Longrightarrow \quad N\left(\mathcal{X}\right) = \left\{\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}\right\}. \quad \Longrightarrow \quad \dim\left\{N\left(\mathcal{X}\right)\right\} = \operatorname{rank}\left(\mathcal{X}\right) - 3 = 0$$

- 设 $\mathcal{X} = \begin{pmatrix} 2 & 3 & 1 \\ 4 & 6 & 2 \\ 6 & 8 & 3 \\ 8 & 2 & 4 \end{pmatrix}$.

$$\Longrightarrow \quad \operatorname{rank}\left(\mathcal{X}\right) = 2. \quad \Longrightarrow \quad \dim\left\{C\left(\mathcal{X}\right)\right\} = 2.$$

$$\Longrightarrow \dim\left\{N\left(\mathcal{X}\right)\right\} = 3 - \operatorname{rank}\left(\mathcal{X}\right) = 1$$

几何直观 (geometrical aspects)

- 投影矩阵 (projection matrix)
  - $\mathbb{R}^n$ 中的一个矩阵 $\mathscr{P}_{n\times n}$ 称为**投影矩阵**，当且仅当

$$\mathscr{P}=\mathscr{P}^{\mathrm{T}}=\mathscr{P}^2 \quad \left(\mathscr{P}\text{ 是对称幂等矩阵}\right)$$

  - 设 $\boldsymbol{b}\in\mathbb{R}^n$，则 $\boldsymbol{b}$ 在 $C\left(\mathscr{P}\right)$ 上的投影为 $\boldsymbol{a}=\mathscr{P}\boldsymbol{b}$.

几何直观 (geometrical aspects)

- 在 $C(\mathcal{X})$ 上的投影
  - 考虑矩阵 $\mathcal{X}_{n\times p}$，记

$\mathcal{P}$ 是对称矩阵：$\mathcal{P}^{\mathrm{T}}=\mathcal{P}$　　$\mathcal{Q}$ 也是对称矩阵：$\mathcal{Q}^{\mathrm{T}}=\mathcal{Q}$

$$\mathcal{P}=\mathcal{X}\left(\mathcal{X}^{\mathrm{T}}\mathcal{X}\right)^{-1}\mathcal{X}^{\mathrm{T}},\quad \mathcal{Q}=\mathcal{I}_n-\mathcal{P}$$

$$\Longrightarrow\quad \mathcal{P}^2=\left[\mathcal{X}\left(\mathcal{X}^{\mathrm{T}}\mathcal{X}\right)^{-1}\mathcal{X}^{\mathrm{T}}\right]\mathcal{X}\left(\mathcal{X}^{\mathrm{T}}\mathcal{X}\right)^{-1}\mathcal{X}^{\mathrm{T}}$$

$$=\mathcal{X}\left(\mathcal{X}^{\mathrm{T}}\mathcal{X}\right)^{-1}\mathcal{X}^{\mathrm{T}}=\mathcal{P}\ \Longrightarrow\ \mathcal{P}\text{ 是幂等阵}$$

$$\Longrightarrow\quad \mathcal{Q}^2=\left(\mathcal{I}_n-\mathcal{P}\right)\left(\mathcal{I}_n-\mathcal{P}\right)$$

$$=\mathcal{I}_n-\mathcal{P}-\mathcal{P}+\mathcal{P}^2=\mathcal{I}_n-\mathcal{P}=\mathcal{Q}\ \Longrightarrow\ \mathcal{Q}\text{ 也是幂等阵}$$

几何直观 (geometrical aspects)

- 在 $C(\mathscr{X})$ 上的投影
  - 考虑矩阵 $\mathscr{X}_{n\times p}$，记

$$\mathscr{P}=\mathscr{X}\left(\mathscr{X}^{\mathrm{T}}\mathscr{X}\right)^{-1}\mathscr{X}^{\mathrm{T}},\quad \mathscr{Q}=\mathscr{I}_n-\mathscr{P}$$

$$\Longrightarrow\quad \mathscr{P}\mathscr{X}=\mathscr{X}\left(\mathscr{X}^{\mathrm{T}}\mathscr{X}\right)^{-1}\mathscr{X}^{\mathrm{T}}\cdot\mathscr{X}=\mathscr{X}$$

  - 由于 $\mathscr{X}$ 的列向量到自身的投影就是它自己，所以对任何向量 $\boldsymbol{b}\in\mathbb{R}^n$，投影矩阵 $\mathscr{P}$ 将它投影到了 $C(\mathscr{X})$ 上.

$$\Longrightarrow\quad \mathscr{Q}\mathscr{X}=\left(\mathscr{I}_n-\mathscr{P}\right)\mathscr{X}=\mathscr{X}-\mathscr{P}\mathscr{X}=0$$

  - 类似地，投影矩阵 $\mathscr{Q}$ 将 $\mathbb{R}^n$ 中的任一向量 $\boldsymbol{b}\in\mathbb{R}^n$ 投影到了 $C(\mathscr{X})$ 的正交补上.

几何直观 (geometrical aspects)

- 定理 2.8 设 $\mathscr{P}=\mathscr{X}\left(\mathscr{X}^{\mathrm{T}}\mathscr{X}\right)^{-1}\mathscr{X}^{\mathrm{T}}$ 为投影矩阵，$\mathscr{Q}=\mathscr{I}_n-\mathscr{P}$ 是其正交补，则

$$(i)\ \boldsymbol{x}=\mathscr{P}\boldsymbol{b} \quad\Longrightarrow\quad \boldsymbol{x}\in C\left(\mathscr{X}\right).$$

$$(ii)\ \boldsymbol{y}=\mathscr{Q}\boldsymbol{b} \quad\Longrightarrow\quad \forall \boldsymbol{x}\in C\left(\mathscr{X}\right),\ \boldsymbol{y}^{\mathrm{T}}\boldsymbol{x}=0.$$

证明 *(i)*

$$\boldsymbol{x}=\mathscr{P}\boldsymbol{b}=\mathscr{X}\left(\mathscr{X}^{\mathrm{T}}\mathscr{X}\right)^{-1}\mathscr{X}^{\mathrm{T}}\boldsymbol{b}$$

$$\boldsymbol{a}=\left(\mathscr{X}^{\mathrm{T}}\mathscr{X}\right)^{-1}\mathscr{X}^{\mathrm{T}}\boldsymbol{b}\in\mathbb{R}^{p}$$

$$=\mathscr{X}\boldsymbol{a}$$

$$=\begin{pmatrix}\boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_p\end{pmatrix}\begin{pmatrix}a_1\\ a_2\\ \vdots\\ a_p\end{pmatrix}$$

$$=a_1\boldsymbol{x}_1+a_2\boldsymbol{x}_2+\cdots+a_p\boldsymbol{x}_p\ \in C\left(\mathscr{X}\right)$$

几何直观 (geometrical aspects)

- 定理 2.8 设 $\mathcal{P}=\mathcal{X}\left(\mathcal{X}^{\mathrm{T}}\mathcal{X}\right)^{-1}\mathcal{X}^{\mathrm{T}}$ 为投影矩阵，$\mathcal{Q}=\mathcal{I}_n-\mathcal{P}$ 是其正交补，则

$$(i)\ \boldsymbol{x}=\mathcal{P}\boldsymbol{b} \implies \boldsymbol{x}\in C\left(\mathcal{X}\right).$$

$$(ii)\ \boldsymbol{y}=\mathcal{Q}\boldsymbol{b} \implies \forall \boldsymbol{x}\in C\left(\mathcal{X}\right),\ \boldsymbol{y}^{\mathrm{T}}\boldsymbol{x}=0.$$

证明 *(ii)*

$$\boldsymbol{y}^{\mathrm{T}}\boldsymbol{x}=\left(\mathcal{Q}\boldsymbol{b}\right)^{\mathrm{T}}\mathcal{X}\boldsymbol{a}=\boldsymbol{b}^{\mathrm{T}}\left(\mathcal{I}_n-\mathcal{P}\right)^{\mathrm{T}}\mathcal{X}\boldsymbol{a}$$

$$=\boldsymbol{b}^{\mathrm{T}}\left(\mathcal{I}_n-\mathcal{P}\right)\mathcal{X}\boldsymbol{a}$$

$$=\boldsymbol{b}^{\mathrm{T}}\mathcal{X}\boldsymbol{a}-\boldsymbol{b}^{\mathrm{T}}\mathcal{X}\left(\mathcal{X}^{\mathrm{T}}\mathcal{X}\right)^{-1}\mathcal{X}^{\mathrm{T}}\mathcal{X}\boldsymbol{a}$$

$\left(\mathcal{X}^{\mathrm{T}}\mathcal{X}\right)^{-1}\mathcal{X}^{\mathrm{T}}\mathcal{X}=\mathcal{I}_p$

$$=\boldsymbol{b}^{\mathrm{T}}\mathcal{X}\boldsymbol{a}-\boldsymbol{b}^{\mathrm{T}}\mathcal{X}\boldsymbol{a}=0$$

几何直观 (geometrical aspects)

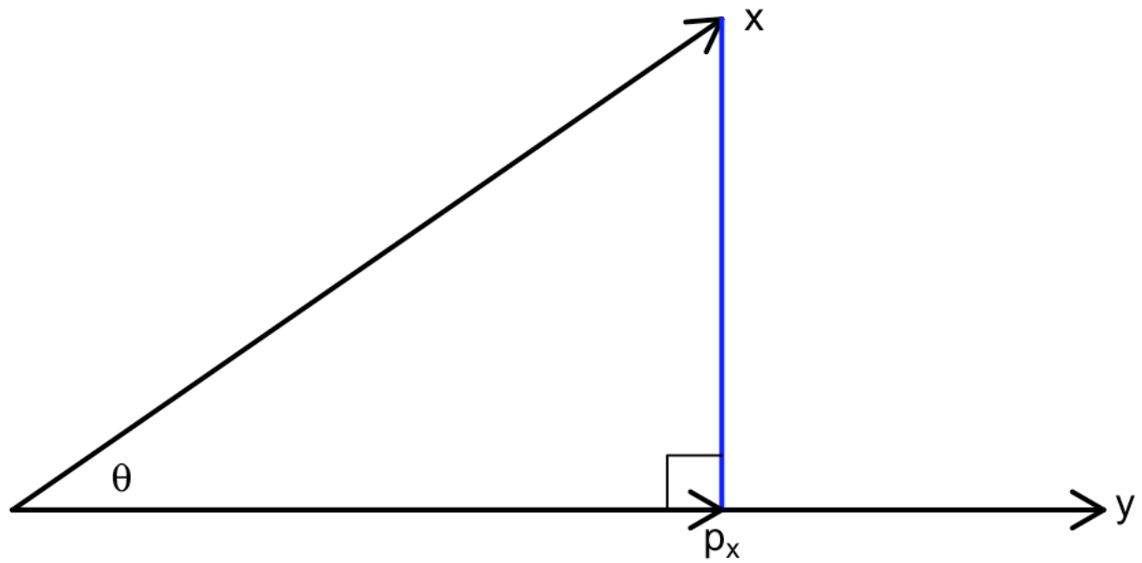
- 注 2.3 设 $x, y \in \mathbb{R}^n$，$p_x \in \mathbb{R}^n$ 表示 x 在 y 上的投影向量.

取 $\mathcal{X} = y$ 时，我们有

$$p_x = \underbrace{y\left(y^{\mathrm{T}}y\right)^{-1}y^{\mathrm{T}}}_{\text{投影矩阵}}x$$

$$= \frac{y^{\mathrm{T}}x}{\|y\|^2}y$$

$$\Longrightarrow \quad \|p_x\| = \sqrt{p_x^{\mathrm{T}}p_x} = \frac{\left|y^{\mathrm{T}}x\right|}{\|y\|}$$



向量 x 在 y 上的投影：$p_x = \dfrac{x^{\mathrm{T}}y}{y^{\mathrm{T}}y}y$

第 3 章

多元中的一些概念与模型

概述

第 3 章 多元中的一些概念与模型

协方差

相关系数

描述性统计量

两个变量的线性模型

单因子方差分析

多重线性模型

Boston 住房数据集

协方差 (Covariance)

- 协方差是对随机变量之间相关性的一种度量.
  - 给定两个随机变量 X 与 Y，它们之间的协方差定义为

$$\sigma_{XY} = \mathrm{Cov}(X, Y) = \mathrm{E}\left[\left(X - \mathrm{E}(X)\right)\left(Y - \mathrm{E}(Y)\right)\right]$$

$$= \mathrm{E}(XY) - \mathrm{E}(X) \cdot \mathrm{E}(Y)$$

  - 若 X 与 Y 相互独立，则 $\mathrm{Cov}(X, Y) = 0$. 反之，不一定成立.
  - X 与其自身的协方差就是方差：

$$\sigma_{XX} = \mathrm{Var}(X) = \mathrm{Cov}(X, X)$$

协方差 (Covariance)

- 协方差是对随机变量之间相关性的一种度量.
  - 如果 $\boldsymbol{X}=\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$ 是 p 维随机向量，则其**协方差矩阵** (covariance matrix) 定义为

$$\Sigma=\begin{pmatrix} \sigma_{X_1X_1} & \sigma_{X_1X_2} & \cdots & \sigma_{X_1X_p} \\ \sigma_{X_2X_1} & \sigma_{X_2X_2} & \cdots & \sigma_{X_2X_p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_pX_1} & \sigma_{X_pX_2} & \cdots & \sigma_{X_pX_p} \end{pmatrix}$$

$$\begin{aligned} \sigma_{X_iX_j} &= \operatorname{Cov}\left(X_i, X_j\right) \\ &= \mathrm{E}\left[\left(X_i-\mathrm{E}\left(X_i\right)\right)\left(X_j-\mathrm{E}\left(X_j\right)\right)\right] \\ &= E\left(X_iX_j\right)-E\left(X_i\right)E\left(X_j\right) \end{aligned}$$

协方差 (Covariance)

- 上述概念的样本形式为

$$s_{X_iX_j} = \frac{1}{n}\sum_{k=1}^{n}\left(x_{ki}-\bar{x}_{\cdot i}\right)\left(x_{kj}-\bar{x}_{\cdot j}\right), \quad i, j = 1, 2, \ldots, p$$

样本协方差 → $\hat{\sigma}_{X_iX_j}$

	X_1	X_2	$\cdots$	X_p
Data :	x_{11}	x_{12}	$\cdots$	x_{1p}
	x_{21}	x_{22}	$\cdots$	x_{2p}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	x_{n1}	x_{n2}	$\cdots$	x_{np}
mean :	$\bar{x}_{\cdot 1}$	$\bar{x}_{\cdot 2}$	$\cdots$	$\bar{x}_{\cdot p}$

  - 对于较小的 $n\,(\leq 20)$，我们有

$$s_{X_iX_j} = \frac{1}{n-1}\sum_{k=1}^{n}\left(x_{ki}-\bar{x}_{\cdot i}\right)\left(x_{kj}-\bar{x}_{\cdot j}\right), \quad i, j = 1, 2, \ldots, p$$

  - 样本协方差矩阵为

$$\mathcal{S} = \begin{pmatrix} s_{X_1X_1} & s_{X_1X_2} & \cdots & s_{X_1X_p} \\ s_{X_2X_1} & s_{X_2X_2} & \cdots & s_{X_2X_p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{X_pX_1} & s_{X_pX_2} & \cdots & s_{X_pX_p} \end{pmatrix}$$

协方差 (Covariance)

- 例: 如果 $\mathcal{X}$ 是所有钞票数据集，则其样本协方差矩阵 $\mathcal{S}$ 为

```
library(mclust)
data(banknote)
head(banknote)
S <- cov(banknote[, 2:7])
round(S, digits = 2)
```

	X_1 Length	X_2 Left	X_3 Right	X_4 Bottom	X_5 Top	X_6 Diagonal
Length	0.14	0.03	0.02	-0.10	-0.02	0.08
Left	0.03	0.13	0.11	0.22	0.11	-0.21
Right	0.02	0.11	0.16	0.28	0.13	-0.24
Bottom	-0.10	0.22	0.28	2.09	0.16	-1.04
Top	-0.02	0.11	0.13	0.16	0.64	-0.55
Diagonal	0.08	-0.21	-0.24	-1.04	-0.55	1.33

$\hat{\sigma}_{X_3X_5} = s_{X_3X_5} = 0.13$

$\widehat{\mathrm{Var}}(X_i) = s_{X_iX_i}$, $i = 1, 2, \ldots, 6$

$\hat{\sigma}_{X_2X_6} = s_{X_2X_6} = -0.21$

协方差 (Covariance)

- 例: 如果 $\mathcal{X}$ 是所有钞票数据集，则其样本协方差矩阵 $\mathcal{S}$ 为

```
library(ggplot2)
fig_1 <- ggplot(data = banknote, aes(x = Left, y = Diagonal, color = Status, size = 1.5)) +
  geom_point()
fig_1
```

$$s_{X_2X_6} = -0.21$$

协方差 (Covariance)

- 例: 如果 $\mathcal{X}$ 是所有钞票数据集，则其样本协方差矩阵 $\mathcal{S}$ 为

```
fig_2 <- ggplot(data = banknote, aes(x = Right, y = Top, color = Status, size = 1.5)) +
  geom_point()
fig_2
```

$s_{X_3X_5} = 0.13$

协方差 (Covariance)

- 例: 如果 $\mathcal{X}_f$ 代表伪钞的数据子集，则其协方差矩阵 $\mathcal{S}_f$ 为

```
counterfeit <- subset(banknote, Status == "counterfeit")
S_f <- cov(counterfeit[, 2:7])
round(S_f, digits = 3)
```

	X_1	X_2	X_3	X_4	X_5	X_6
	Length	Left	Right	Bottom	Top	Diagonal
Length	0.124	0.032	0.024	-0.101	0.019	0.012
Left	0.032	0.065	0.047	-0.024	-0.012	-0.005
Right	0.024	0.047	0.089	-0.019	0.000	0.034
Bottom	-0.101	-0.024	-0.019	1.281	-0.490	0.238
Top	0.019	-0.012	0.000	-0.490	0.404	-0.022
Diagonal	0.012	-0.005	0.034	0.238	-0.022	0.311

$\hat{\sigma}_{X_3X_5} = s_{X_3X_5} = 0.000$

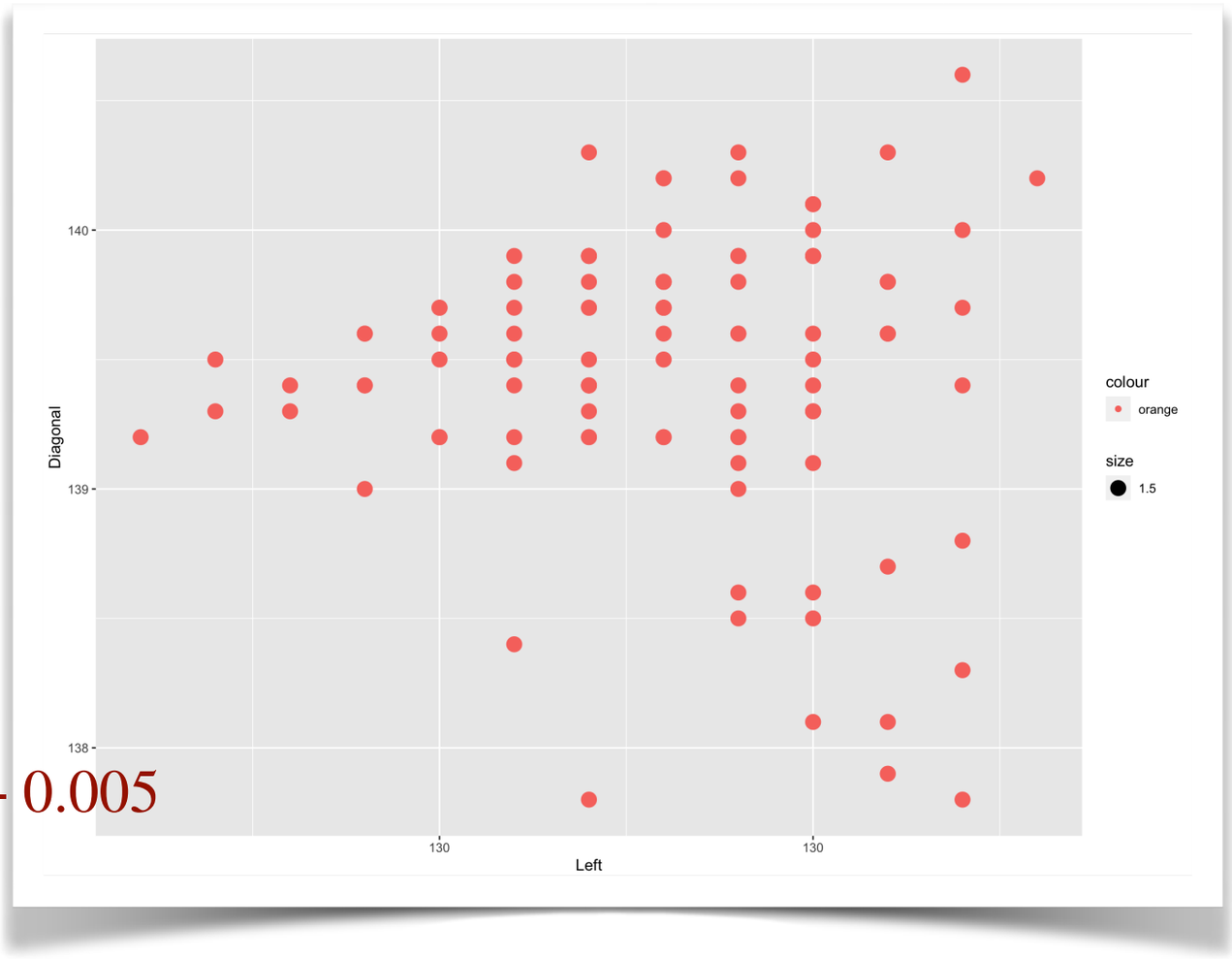
$\widehat{\mathrm{Var}}(X_i) = s_{X_iX_i}$, $i = 1, 2, \ldots, 6$

$\hat{\sigma}_{X_2X_6} = s_{X_2X_6} = -0.005$

协方差 (Covariance)

- 例: 如果 $\mathcal{X}_f$ 代表伪钞的数据子集，则其协方差矩阵 $\mathcal{S}_f$ 为

```
fig_3 <- ggplot(data = counterfeit, aes(x = Left, y = Diagonal, size = 1.5, color = 'orange')) +
  geom_point()
fig_3
```

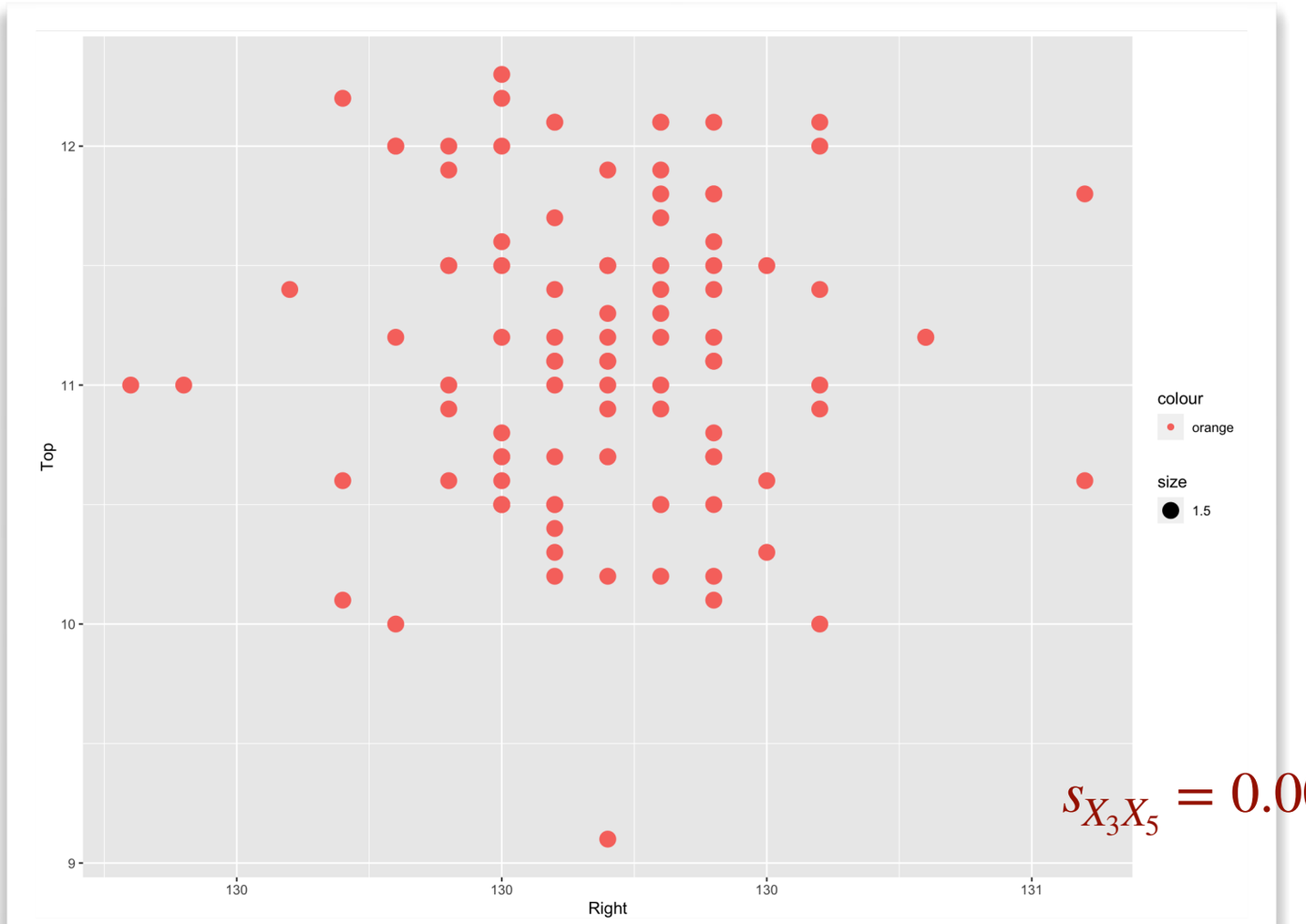


$$s_{X_2X_6} = -0.005$$

协方差 (Covariance)

- 例: 如果 $\mathcal{X}_f$ 代表伪钞的数据子集，则其协方差矩阵 $\mathcal{S}_f$ 为

```
fig_4 <- ggplot(data = counterfeit, aes(x = Right, y = Top, size = 1.5, color = 'orange')) +
  geom_point()
fig_4
```



$s_{X_3X_5} = 0.000$

协方差 (Covariance)

- 例: 如果 $\mathcal{X}_g$ 表示真钞的数据子集，则其协方差矩阵 $\mathcal{S}_g$ 为

```
counterfeit <- subset(banknote, Status == "counterfeit")
S_f <- cov(counterfeit[, 2:7])
round(S_f, digits = 3)
```

	X_1	X_2	X_3	X_4	X_5	X_6
	Length	Left	Right	Bottom	Top	Diagonal
Length	0.150	0.058	0.057	0.057	0.014	0.005
Left	0.058	0.133	0.086	0.057	0.049	-0.043
Right	0.057	0.086	0.126	0.058	0.031	-0.024
Bottom	0.057	0.057	0.058	0.413	-0.263	0.000
Top	0.014	0.049	0.031	-0.263	0.421	-0.075
Diagonal	0.005	-0.043	-0.024	0.000	-0.075	0.200

$\hat{\sigma}_{X_3X_5} = s_{X_3X_5} = 0.031$

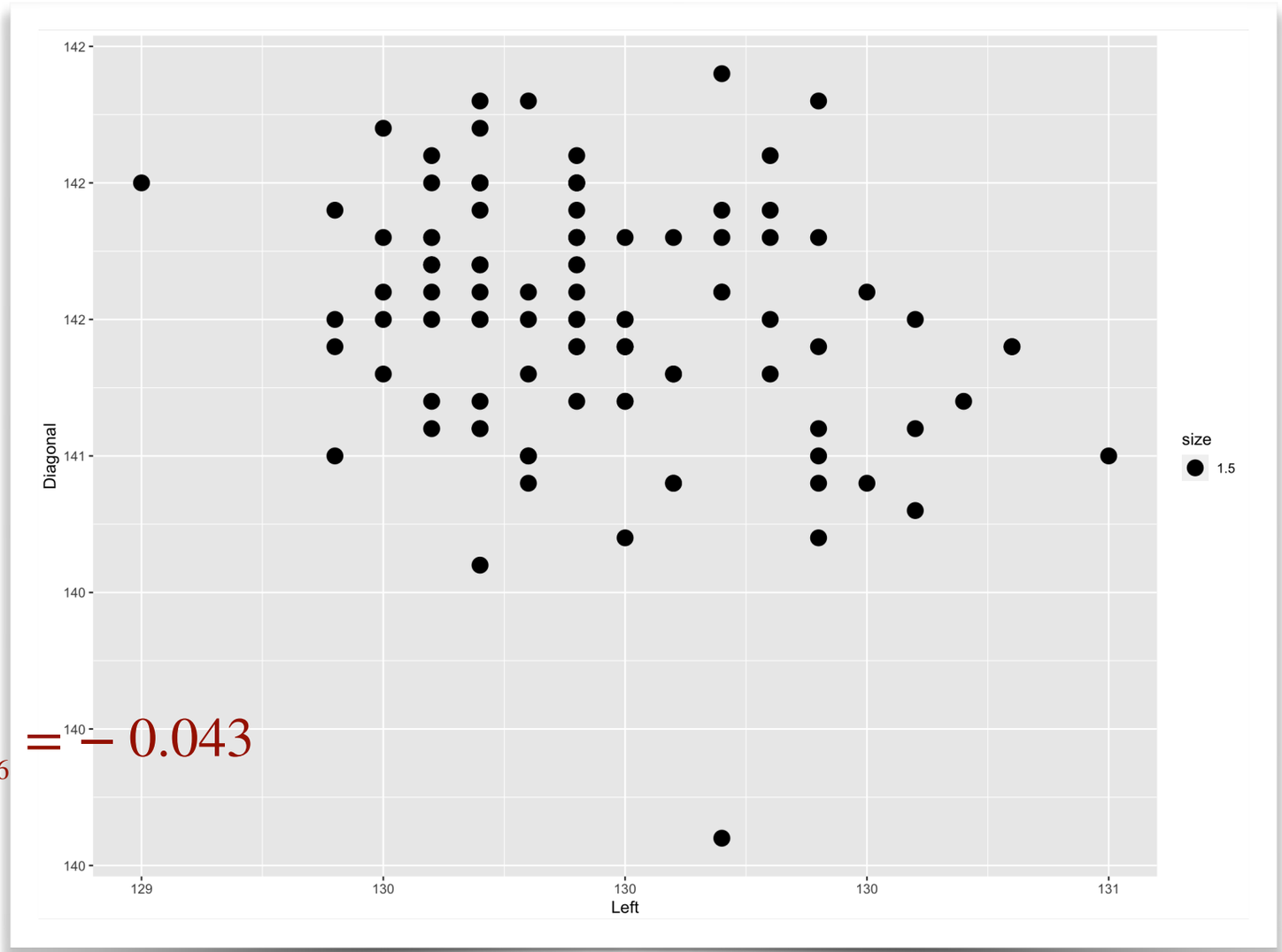
$\widehat{\mathrm{Var}}(X_i) = s_{X_iX_i}$, $i = 1, 2, \ldots, 6$

$\hat{\sigma}_{X_2X_6} = s_{X_2X_6} = -0.043$

协方差 (Covariance)

- 例: 如果 $\mathcal{X}_g$ 表示真钞的数据子集，则其协方差矩阵 $\mathcal{S}_g$ 为

```
fig_5 <- ggplot(data = genuine, aes(x = Left, y = Diagonal, size = 1.5)) +
  geom_point()
fig_5
```

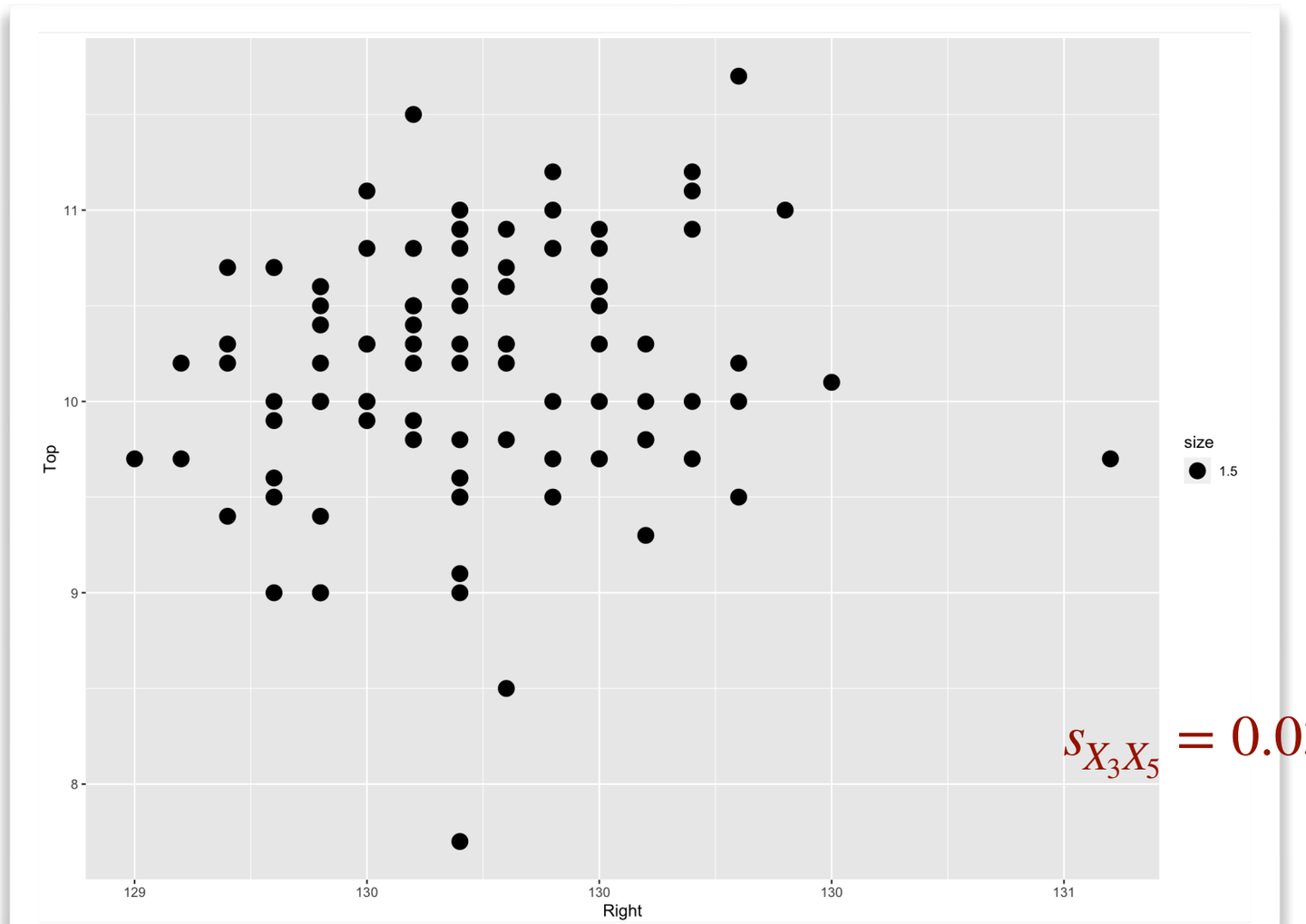


$s_{X_2X_6} = -0.043$

协方差 (Covariance)

- 例: 如果 $\mathcal{X}_g$ 表示真钞的数据子集，则其协方差矩阵 $\mathcal{S}_g$ 为

```
fig_6 <- ggplot(data = genuine, aes(x = Right, y = Top, size = 1.5,)) +
  geom_point()
fig_6
```



$s_{X_3X_5} = 0.031$

相关系数 (Correlation)

- 两个随机变量 X 与 Y 的**相关系数** (correlation coefficient) 可由它们的协方差定义如下:

$$\rho_{XY} = \frac{\mathrm{Cov}(X,\ Y)}{\sqrt{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)}}$$

  - 相关系数的优点在于它和变量的量纲无关.
  - $\left|\rho_{XY}\right| \leq 1.$
  - $\left|\rho_{XY}\right| = 1$ 当且仅当 $P\left(Y = aX + b\right) = 1.$

相关系数 (Correlation)

- 对 p 维随机向量 $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$，我们有如下的**相关矩阵** (correlation matrix)

$$\mathscr{P} = \begin{pmatrix} \rho_{X_1X_1} & \rho_{X_1X_2} & \cdots & \rho_{X_1X_p} \\ \rho_{X_2X_1} & \rho_{X_2X_2} & \cdots & \rho_{X_2X_p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{X_pX_1} & \rho_{X_pX_2} & \cdots & \rho_{X_pX_p} \end{pmatrix} = \begin{pmatrix} 1 & \rho_{X_1X_2} & \cdots & \rho_{X_1X_p} \\ \rho_{X_2X_1} & 1 & \cdots & \rho_{X_2X_p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{X_pX_1} & \rho_{X_pX_2} & \cdots & 1 \end{pmatrix}$$

相关系数 (Correlation)

- 上述概念的样本形式为

$$s_{X_iX_j} = \frac{1}{n}\sum_{k=1}^{n}\left(x_{ki}-\bar{x}_{\cdot i}\right)\left(x_{kj}-\bar{x}_{\cdot j}\right), \quad i, j = 1, 2, \ldots, p$$

$$r_{X_iX_j} = \frac{s_{X_iX_j}}{\sqrt{s_{X_iX_i}\cdot s_{X_jX_j}}}, \quad i, j = 1, 2, \ldots, p$$

$r_{X_iX_j}$：X_i 与 X_j 的样本相关系数

$$\text{Data}: \begin{matrix} X_1 & X_2 & \cdots & X_p \\ x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \\ \bar{x}_{\cdot 1} & \bar{x}_{\cdot 2} & \cdots & \bar{x}_{\cdot p} \end{matrix} \quad \text{mean}:$$

- 对较小的 $n\,(\le 20)$，我们有

$$s_{X_iX_j} = \frac{1}{n-1}\sum_{k=1}^{n}\left(x_{ki}-\bar{x}_{\cdot i}\right)\left(x_{kj}-\bar{x}_{\cdot j}\right), \quad i, j = 1, 2, \ldots, p$$

- 样本相关矩阵为

$$\mathscr{R} = \begin{pmatrix} r_{X_1X_1} & r_{X_1X_2} & \cdots & r_{X_1X_p} \\ r_{X_2X_1} & r_{X_2X_2} & \cdots & r_{X_2X_p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{X_pX_1} & r_{X_pX_2} & \cdots & r_{X_pX_p} \end{pmatrix} = \begin{pmatrix} 1 & r_{X_1X_2} & \cdots & r_{X_1X_p} \\ r_{X_2X_1} & 1 & \cdots & r_{X_2X_p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{X_pX_1} & r_{X_pX_2} & \cdots & 1 \end{pmatrix}$$

相关系数 (Correlation)

- 例: 对真钞数据子集，可得其样本相关矩阵如下

```
R_g <- cor(genuine[, 2:7])
round(R_g, digits = 2)
```

	X_1	X_2	X_3	X_4	X_5	X_6
	Length	Left	Right	Bottom	Top	Diagonal
Length	1.00	0.41	0.42	0.23	0.06	0.03
Left	0.41	1.00	0.66	0.24	0.21	-0.26
Right	0.42	0.66	1.00	0.25	0.13	-0.15
Bottom	0.23	0.24	0.25	1.00	-0.63	0.00
Top	0.06	0.21	0.13	-0.63	1.00	-0.26
Diagonal	0.03	-0.26	-0.15	0.00	-0.26	1.00

相关系数 (Correlation)

- 例: 对伪钞数据子集，可得其样本相关矩阵如下

```
R_f <- cor(counterfeit[, 2:7])
round(R_f, digits = 2)
```

	X_1	X_2	X_3	X_4	X_5	X_6
	Length	Left	Right	Bottom	Top	Diagonal
Length	1.00	0.35	0.23	-0.25	0.09	0.06
Left	0.35	1.00	0.61	-0.08	-0.07	-0.04
Right	0.23	0.61	1.00	-0.06	0.00	0.21
Bottom	-0.25	-0.08	-0.06	1.00	-0.68	0.38
Top	0.09	-0.07	0.00	-0.68	1.00	-0.06
Diagonal	0.06	-0.04	0.21	0.38	-0.06	1.00

相关系数 (Correlation)

定理 3.1 如果 X 与 Y 相互独立，则 $\rho(X, Y) = \mathrm{Cov}(X, Y) = 0.$

一般情形下，其逆命题不成立.

- 例: $X \sim N(0, 1)$, $Y = X^2$. 显然，X 与 Y 不独立.

$$\begin{aligned}\mathrm{Cov}(X, Y) &= \mathrm{E}(XY) - \mathrm{E}(X) \cdot \mathrm{E}(Y) \\ &= \mathrm{E}\left(X^3\right) - 0 \cdot \mathrm{E}(Y) = 0\end{aligned}$$

相关系数 (Correlation)

注 3.1 对于两个服从正态分布的随机变量而言，定理3.1的逆命题成立：两个服从正态分布的随机变量不相关也意味着它们相互独立.

- 定理3.1 和 注3.1 可用于确定二元正态分布的两个分量的独立性.
- 不幸的是，对任意的二元分布 (X, Y) 而言，r_{XY} 的分布则比较复杂.
- Fisher 的 Z 变换：$W = \frac{1}{2}\log\left(\frac{1+r_{XY}}{1-r_{XY}}\right)$.

可以证明 $\Longrightarrow$

$$\mathrm{E}(W) \approx \frac{1}{2}\log\left(\frac{1+\rho_{XY}}{1-\rho_{XY}}\right)$$

$$\mathrm{Var}(W) \approx \frac{1}{n-3}$$

相关系数 (Correlation)

定理 3.2

$$Z = \frac{W - \mathrm{E}(W)}{\sqrt{\mathrm{Var}(W)}} \xrightarrow{\mathscr{L}} N(0,\ 1)$$

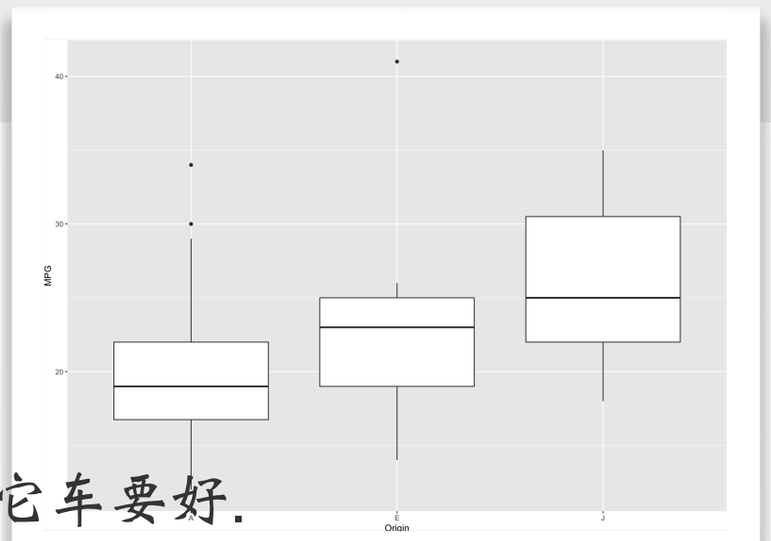
依分布收敛

- 例: 汽车数据集当中， 变量 MPG 与 Weight 的相关系数.

```
library(corrgram)
str(auto)
head(auto)
cor(auto$MPG, auto$Weight)
```

```
> cor(auto$MPG, auto$Weight)
[1] -0.8228964
```

```
library(ggplot2)
ggplot(auto, aes(x = Origin, y = MPG)) + geom_boxplot()
```



- 结论： 日本车 (每加仑汽油) 的里程数一般比其它车要好.

相关系数 (Correlation)

定理 3.2

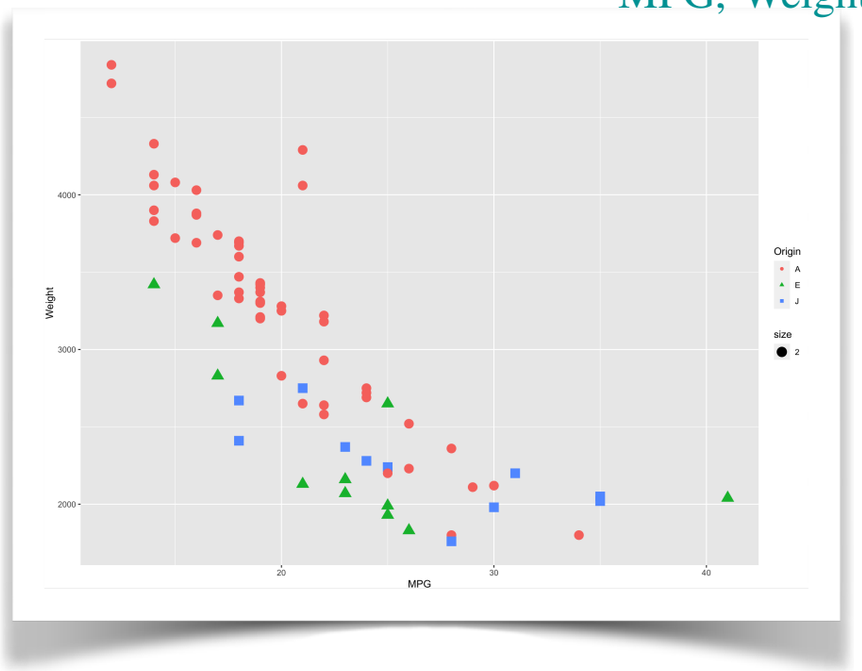
$$Z = \frac{W - \mathrm{E}(W)}{\sqrt{\mathrm{Var}(W)}} \xrightarrow{\mathscr{L}} N(0,\ 1)$$

依分布收敛

- 例: 汽车数据集当中， 变量 MPG 与 Weight 的相关系数.

```
ggplot(auto, aes(x = MPG, y = Weight, color = Origin, shape = Origin, size = 2)) +
geom_point()
```

$$r_{\mathrm{MPG},\ \mathrm{Weight}} = -0.823$$



- 日本汽车比其他汽车轻.

相关系数 (Correlation)

定理 3.2

$$Z = \frac{W - \mathrm{E}(W)}{\sqrt{\mathrm{Var}(W)}} \xrightarrow{\mathscr{L}} N(0,\ 1)$$

依分布收敛

- 例: 汽车数据集当中， 变量 MPG 与 Weight 的相关系数.
  - 我们可以应用 Fisher 的 Z 变换来检验 $\rho_{\mathrm{MPG,\ Weight}}$ 是否与 $\rho_0 = 0$ 有显著差异.

```
w <- log((1 + cor(auto$MPG, auto$Weight)) / (1 - cor(auto$MPG, auto$Weight))) / 2
w
```

```
> w
[1] -1.165723
```

```
z <- (w - 0) / sqrt(1/(74-3))
z
```

```
> z
[1] -9.82256
```

$$w = \frac{1}{2}\log\left(\frac{1 + r_{\mathrm{MPG,\ Weight}}}{1 - r_{\mathrm{MPG,\ Weight}}}\right) = -1.165723\ ,\quad z = \frac{-1.165723 - 0}{\sqrt{\frac{1}{74 - 3}}} = -9.82256$$

相关系数 (Correlation)

定理 3.2

$$Z = \frac{W - \mathrm{E}(W)}{\sqrt{\mathrm{Var}(W)}} \xrightarrow{\mathscr{L}} N(0,\ 1)$$

依分布收敛

- 例: 汽车数据集当中， 变量 MPG 与 Weight 的相关系数.
  - $\mathrm{H}_0:\ \rho_{\mathrm{MPG,\ Weight}} = 0 \longleftrightarrow \mathrm{H}_1:\ \rho_{\mathrm{MPG,\ Weight}} \neq 0$

```
p_value <- 2 * pnorm(abs(z), 0, 1, lower.tail = FALSE)
p_value
```

```
> p_value
[1] 9.002782e-23
```

$\Longrightarrow$ Reject $\mathrm{H}_0:\ \rho_{\mathrm{MPG,\ Weight}} = 0$

$$w = \frac{1}{2}\log\left(\frac{1 + r_{\mathrm{MPG,\ Weight}}}{1 - r_{\mathrm{MPG,\ Weight}}}\right) = -1.165723\ ,\quad z = \frac{-1.165723 - 0}{\sqrt{\frac{1}{74 - 3}}} = -9.82256$$

相关系数 (Correlation)

定理 3.2

$$Z = \frac{W - \mathrm{E}(W)}{\sqrt{\mathrm{Var}(W)}} \xrightarrow{\mathscr{L}} N(0,\ 1)$$

依分布收敛

- 例: 汽车数据集当中， 变量 MPG 与 Weight 的相关系数.
  - $\mathrm{H}_0:\ \rho_{\mathrm{MPG,\ Weight}} = -0.75 \longleftrightarrow \mathrm{H}_1:\ \rho_{\mathrm{MPG,\ Weight}} \neq -0.75$

```
E_w <- log((1 + (-0.75)) / (1 - (-0.75))) / 2
E_w
```

```
> E_w
[1] -0.9729551
```

```
z1 <- (w - (E_w)) / sqrt(1/(74-3))
z1
```

```
> z1
[1] -1.624295
```

```
p_value <- 2 * pnorm(abs(z1), 0, 1, lower.tail = FALSE)
p_value
```

```
> p_value
[1] 0.1043129
```

$> 0.05 \Longrightarrow$ Accept $\mathrm{H}_0:\ \rho_{\mathrm{MPG,\ Weight}} = -0.75$

相关系数 (Correlation)

注 3.2 W 正态分布以及方差稳定的性质是渐进性质. 此外, 在小样本 $\left(n \leq 25\right)$ 时, 往往对 W 作改进后的 Hotelling 变换

$$W^* = W - \frac{3W + \tanh(W)}{4(n-1)}, \quad \text{且 } \operatorname{Var}(W^*) = \frac{1}{n-1}$$

变换后的变量 W^* 渐近服从正态分布.

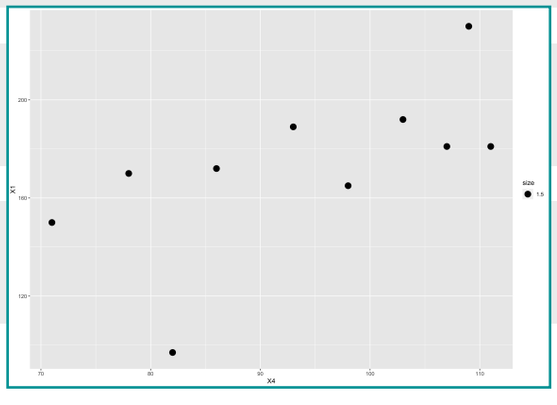
相关系数 (Correlation)

- 例: 某纺织品店经理研究“经典蓝色”的套头衫在10个不同时期的销售情况. 他收集了套头衫的销售量 (X_1)，价格的变化 (X_2) (单位：欧元)，在当地报纸的广告费用 (X_3) (单位：欧元)，以及促销员的工作时长 (X_4) (每个时段的小时数). 在这些时期内，他观测到的数据矩阵如下：

```
x <- c(230, 125, 200, 109, 181, 99, 55, 107, 165, 97, 105, 98, 150, 115, 85, 71, 97, 120, 0, 82,
  192, 100, 150, 103, 181, 80, 85, 111, 189, 90, 120, 93, 172, 95, 110, 86, 170, 125, 130, 78)
pullover <- matrix(x, ncol = 4, byrow = TRUE)
pullover <- as.data.frame(pullover)
colnames(pullover) <- c("X1", "X2", "X3", "X4")
pullover
```

```
> pullover
    X1  X2  X3  X4
1  230 125 200 109
2  181  99  55 107
3  165  97 105  98
4  150 115  85  71
5   97 120   0  82
6  192 100 150 103
7  181  80  85 111
8  189  90 120  93
9  172  95 110  86
10 170 125 130  78
```

```
ggplot(pullover, aes(x = X4, y = X1, size = 1.5)) +
  geom_point()
```



```
r_14 <- cor(pullover$V1, pullover$V4)
r_14
```

```
> r_14
[1] 0.6328673
```

相关系数 (Correlation)

- 例: 某纺织品店经理研究“经典蓝色”的套头衫在10个不同时期的销售情况. 他收集了套头衫的销售量 (X_1)，价格的变化 (X_2) (单位：欧元)，在当地报纸的广告费用 (X_3) (单位：欧元)，以及促销员的工作时长 (X_4) (每个时段的小时数). 在这些时期内，他观测到的数据矩阵如下：

```
w <- 0.5 * (log((1 + r_14)/(1 - r_14)))
w
```

```
> w
[1] 0.7461847
```

```
w_star <- w - (3 * w + tanh(w)) / (4 * (10 - 1))
w_star
```

```
> w_star
[1] 0.666423
```

- $H_0: \rho_{X_1X_4} = 0 \longleftrightarrow H_1: \rho_{X_1X_4} \neq 0$

```
z <- (w_star - 0) / (sqrt(1/(10-1)))
z
```

```
> z
[1] 1.999269
```

```
p_value <- 2 * pnorm(abs(z), 0, 1, lower.tail = FALSE)
p_value
```

```
> p_value
[1] 0.04557925
```

< 0.05 拒绝 $H_0: \rho_{X_1X_4} = 0$

```
> pullover
   X1  X2  X3  X4
1  230 125 200 109
2  181  99  55 107
3  165  97 105  98
4  150 115  85  71
5   97 120   0  82
6  192 100 150 103
7  181  80  85 111
8  189  90 120  93
9  172  95 110  86
10 170 125 130  78
```

相关系数 (Correlation)

注 3.3 注意到 Fisher 的 Z 变换是双曲正切函数的逆: $W=\tanh^{-1}\left(r_{XY}\right)$; 等价地有

$$r_{XY}=\tanh(W)=\frac{\mathrm{e}^{2W}-1}{\mathrm{e}^{2W}+1}.$$

注 3.4 在 X 与 Y 均服从正态分布的假设下, 我们可以利用统计量

$$T=r_{XY}\sqrt{\frac{n-2}{1-r_{XY}^2}} \overset{\rho_{XY}=0}{\sim} t_{n-2}$$

服从 t 分布来检验它们的独立性 $\left(\rho_{XY}=0\right)$. 设犯第I类错误的概率为 α, 则当 $\left|T\right|\geq t_{n-2}\left(1-\frac{\alpha}{2}\right)$ 时我们拒绝零假设 $\rho_{XY}=0$.

描述性统计量 (Summary Statistics)

- 讨论基本的描述性统计量 (均值向量、协方差矩阵和相关矩阵) 的矩阵表示.

  - 随机向量 $\boldsymbol{X}=\begin{pmatrix}X_1\\X_2\\\vdots\\X_p\end{pmatrix}$. 数据矩阵 $\mathcal{X}=\begin{pmatrix}x_{11}&x_{12}&\cdots&x_{1p}\\x_{21}&x_{22}&\cdots&x_{2p}\\\vdots&\vdots&\ddots&\vdots\\x_{n1}&x_{n2}&\cdots&x_{np}\end{pmatrix}=\begin{pmatrix}\boldsymbol{x}_1^{\mathrm{T}}\\\boldsymbol{x}_2^{\mathrm{T}}\\\vdots\\\boldsymbol{x}_n^{\mathrm{T}}\end{pmatrix}$.

  - 重心: $\bar{\boldsymbol{x}}=\begin{pmatrix}\bar{x}_1\\\bar{x}_2\\\vdots\\\bar{x}_p\end{pmatrix}=\frac{1}{n}\begin{pmatrix}x_{11}&x_{21}&\cdots&x_{n1}\\x_{12}&x_{22}&\cdots&x_{n2}\\\vdots&\vdots&\ddots&\vdots\\x_{1p}&x_{2p}&\cdots&x_{np}\end{pmatrix}_{p\times n}\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}_{n\times 1}=\frac{1}{n}\mathcal{X}^{\mathrm{T}}\mathbf{1}_n$ (均值向量)

描述性统计量 (Summary Statistics)

- 样本协方差矩阵:

$$\mathcal{S} = \begin{pmatrix} s_{X_1X_1} & s_{X_1X_2} & \cdots & s_{X_1X_p} \\ s_{X_2X_1} & s_{X_2X_2} & \cdots & s_{X_2X_p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{X_pX_1} & s_{X_pX_2} & \cdots & s_{X_pX_p} \end{pmatrix}$$

Data matrix: $\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \boldsymbol{x}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{pmatrix}$

$$= \frac{1}{n} \begin{pmatrix} \sum_{k=1}^{n} (x_{k1} - \bar{x}_1)^2 & \sum_{k=1}^{n} (x_{k1} - \bar{x}_1)(x_{k2} - \bar{x}_2) & \cdots & \sum_{k=1}^{n} (x_{k1} - \bar{x}_1)(x_{kp} - \bar{x}_p) \\ \sum_{k=1}^{n} (x_{k2} - \bar{x}_2)(x_{k1} - \bar{x}_1) & \sum_{k=1}^{n} (x_{k2} - \bar{x}_2)^2 & \cdots & \sum_{k=1}^{n} (x_{k2} - \bar{x}_2)(x_{kp} - \bar{x}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^{n} (x_{kp} - \bar{x}_p)(x_{k1} - \bar{x}_1) & \sum_{k=1}^{n} (x_{kp} - \bar{x}_p)(x_{k2} - \bar{x}_2) & \cdots & \sum_{k=1}^{n} (x_{kp} - \bar{x}_p)^2 \end{pmatrix}$$

$$= \frac{1}{n} \begin{pmatrix} \sum_{k=1}^{n} (x_{k1} - \bar{x}_1)(x_{k1} - \bar{x}_1) & \sum_{k=1}^{n} (x_{k1} - \bar{x}_1)(x_{k2} - \bar{x}_2) & \cdots & \sum_{k=1}^{n} (x_{k1} - \bar{x}_1)(x_{kp} - \bar{x}_p) \\ \sum_{k=1}^{n} (x_{k2} - \bar{x}_2)(x_{k1} - \bar{x}_1) & \sum_{k=1}^{n} (x_{k2} - \bar{x}_2)(x_{k2} - \bar{x}_2) & \cdots & \sum_{k=1}^{n} (x_{k2} - \bar{x}_2)(x_{kp} - \bar{x}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^{n} (x_{kp} - \bar{x}_p)(x_{k1} - \bar{x}_1) & \sum_{k=1}^{n} (x_{kp} - \bar{x}_p)(x_{k2} - \bar{x}_2) & \cdots & \sum_{k=1}^{n} (x_{kp} - \bar{x}_p)(x_{kp} - \bar{x}_p) \end{pmatrix}$$

描述性统计量 (Summary Statistics)

- 样本协方差矩阵:

$$\mathcal{S} = \begin{pmatrix} s_{X_1X_1} & s_{X_1X_2} & \cdots & s_{X_1X_p} \\ s_{X_2X_1} & s_{X_2X_2} & \cdots & s_{X_2X_p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{X_pX_1} & s_{X_pX_2} & \cdots & s_{X_pX_p} \end{pmatrix}$$

Data matrix: $\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \boldsymbol{x}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{pmatrix}$

$$= \frac{1}{n}\sum_{k=1}^{n} \begin{pmatrix} (x_{k1}-\bar{x}_1)(x_{k1}-\bar{x}_1) & (x_{k1}-\bar{x}_1)(x_{k2}-\bar{x}_2) & \cdots & (x_{k1}-\bar{x}_1)(x_{kp}-\bar{x}_p) \\ (x_{k2}-\bar{x}_2)(x_{k1}-\bar{x}_1) & (x_{k2}-\bar{x}_2)(x_{k2}-\bar{x}_2) & \cdots & (x_{k2}-\bar{x}_2)(x_{kp}-\bar{x}_p) \\ \vdots & \vdots & \ddots & \vdots \\ (x_{kp}-\bar{x}_p)(x_{k1}-\bar{x}_1) & (x_{kp}-\bar{x}_p)(x_{k2}-\bar{x}_2) & \cdots & (x_{kp}-\bar{x}_p)(x_{kp}-\bar{x}_p) \end{pmatrix}$$

$$= \frac{1}{n} \begin{pmatrix} \sum_{k=1}^{n}(x_{k1}-\bar{x}_1)(x_{k1}-\bar{x}_1) & \sum_{k=1}^{n}(x_{k1}-\bar{x}_1)(x_{k2}-\bar{x}_2) & \cdots & \sum_{k=1}^{n}(x_{k1}-\bar{x}_1)(x_{kp}-\bar{x}_p) \\ \sum_{k=1}^{n}(x_{k2}-\bar{x}_2)(x_{k1}-\bar{x}_1) & \sum_{k=1}^{n}(x_{k2}-\bar{x}_2)(x_{k2}-\bar{x}_2) & \cdots & \sum_{k=1}^{n}(x_{k2}-\bar{x}_2)(x_{kp}-\bar{x}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^{n}(x_{kp}-\bar{x}_p)(x_{k1}-\bar{x}_1) & \sum_{k=1}^{n}(x_{kp}-\bar{x}_p)(x_{k2}-\bar{x}_2) & \cdots & \sum_{k=1}^{n}(x_{kp}-\bar{x}_p)(x_{kp}-\bar{x}_p) \end{pmatrix}$$

描述性统计量 (Summary Statistics)

- 样本协方差矩阵:

$$\mathcal{S} = \begin{pmatrix} s_{X_1X_1} & s_{X_1X_2} & \cdots & s_{X_1X_p} \\ s_{X_2X_1} & s_{X_2X_2} & \cdots & s_{X_2X_p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{X_pX_1} & s_{X_pX_2} & \cdots & s_{X_pX_p} \end{pmatrix}$$

Data matrix: $\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \boldsymbol{x}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{pmatrix}$

$$= \frac{1}{n}\sum_{k=1}^{n} \begin{pmatrix} (x_{k1}-\bar{x}_1)(x_{k1}-\bar{x}_1) & (x_{k1}-\bar{x}_1)(x_{k2}-\bar{x}_2) & \cdots & (x_{k1}-\bar{x}_1)(x_{kp}-\bar{x}_p) \\ (x_{k2}-\bar{x}_2)(x_{k1}-\bar{x}_1) & (x_{k2}-\bar{x}_2)(x_{k2}-\bar{x}_2) & \cdots & (x_{k2}-\bar{x}_2)(x_{kp}-\bar{x}_p) \\ \vdots & \vdots & \ddots & \vdots \\ (x_{kp}-\bar{x}_p)(x_{k1}-\bar{x}_1) & (x_{kp}-\bar{x}_p)(x_{k2}-\bar{x}_2) & \cdots & (x_{kp}-\bar{x}_p)(x_{kp}-\bar{x}_p) \end{pmatrix}$$

$$= \frac{1}{n}\sum_{k=1}^{n} \begin{pmatrix} x_{k1}-\bar{x}_1 \\ x_{k2}-\bar{x}_2 \\ \vdots \\ x_{kp}-\bar{x}_p \end{pmatrix} \left(x_{k1}-\bar{x}_1 \, , \, x_{k2}-\bar{x}_2 \, , \, \cdots \, , \, x_{kp}-\bar{x}_p \right)$$

描述性统计量 (Summary Statistics)

- 样本协方差矩阵:

$$\mathcal{S} = \begin{pmatrix} s_{X_1X_1} & s_{X_1X_2} & \cdots & s_{X_1X_p} \\ s_{X_2X_1} & s_{X_2X_2} & \cdots & s_{X_2X_p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{X_pX_1} & s_{X_pX_2} & \cdots & s_{X_pX_p} \end{pmatrix}$$

$$= \frac{1}{n}\sum_{k=1}^{n} \begin{pmatrix} x_{k1} - \bar{x}_1 \\ x_{k2} - \bar{x}_2 \\ \vdots \\ x_{kp} - \bar{x}_p \end{pmatrix} \begin{pmatrix} x_{k1} - \bar{x}_1 \\ x_{k2} - \bar{x}_2 \\ \vdots \\ x_{kp} - \bar{x}_p \end{pmatrix}^{\mathrm{T}}$$

$$= \frac{1}{n}\sum_{k=1}^{n} \begin{pmatrix} x_{k1} - \bar{x}_1 \\ x_{k2} - \bar{x}_2 \\ \vdots \\ x_{kp} - \bar{x}_p \end{pmatrix} \left(x_{k1} - \bar{x}_1 \ , \ x_{k2} - \bar{x}_2 \ , \ \cdots \ , \ x_{kp} - \bar{x}_p \right)$$

Data matrix: $\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \boldsymbol{x}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{pmatrix}$

描述性统计量 (Summary Statistics)

- 样本协方差矩阵:

$$\mathcal{S} = \begin{pmatrix} s_{X_1X_1} & s_{X_1X_2} & \cdots & s_{X_1X_p} \\ s_{X_2X_1} & s_{X_2X_2} & \cdots & s_{X_2X_p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{X_pX_1} & s_{X_pX_2} & \cdots & s_{X_pX_p} \end{pmatrix} = \frac{1}{n}\sum_{k=1}^{n}\left(\boldsymbol{x}_k - \bar{\boldsymbol{x}}\right)\left(\boldsymbol{x}_k - \bar{\boldsymbol{x}}\right)^{\mathrm{T}}$$

Data matrix: $\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \boldsymbol{x}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{pmatrix}$

$$= \frac{1}{n}\sum_{k=1}^{n}\begin{pmatrix} x_{k1} - \bar{x}_1 \\ x_{k2} - \bar{x}_2 \\ \vdots \\ x_{kp} - \bar{x}_p \end{pmatrix}\begin{pmatrix} x_{k1} - \bar{x}_1 \\ x_{k2} - \bar{x}_2 \\ \vdots \\ x_{kp} - \bar{x}_p \end{pmatrix}^{\mathrm{T}} = \frac{1}{n}\sum_{k=1}^{n}\left(\begin{pmatrix} x_{k1} \\ x_{k2} \\ \vdots \\ x_{kp} \end{pmatrix} - \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix}\right)\cdot\left(\begin{pmatrix} x_{k1} \\ x_{k2} \\ \vdots \\ x_{kp} \end{pmatrix} - \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix}\right)^{\mathrm{T}}$$

描述性统计量 (Summary Statistics)

- 样本协方差矩阵：

$$\mathcal{S} = \frac{1}{n}\sum_{k=1}^{n}\left(\boldsymbol{x}_k - \bar{\boldsymbol{x}}\right)\left(\boldsymbol{x}_k - \bar{\boldsymbol{x}}\right)^{\mathrm{T}}$$

Data matrix: $\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \boldsymbol{x}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{pmatrix}$

$$= \frac{1}{n}\sum_{k=1}^{n}\left(\boldsymbol{x}_k - \bar{\boldsymbol{x}}\right)\left(\boldsymbol{x}_k^{\mathrm{T}} - \bar{\boldsymbol{x}}^{\mathrm{T}}\right) = \frac{1}{n}\sum_{k=1}^{n}\left(\boldsymbol{x}_k\boldsymbol{x}_k^{\mathrm{T}} - \boldsymbol{x}_k\bar{\boldsymbol{x}}^{\mathrm{T}} - \bar{\boldsymbol{x}}\boldsymbol{x}_k^{\mathrm{T}} + \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}}\right)$$

$$= \frac{1}{n}\sum_{k=1}^{n}\left(\boldsymbol{x}_k\boldsymbol{x}_k^{\mathrm{T}}\right) - \left(\frac{1}{n}\sum_{k=1}^{n}\boldsymbol{x}_k\right)\bar{\boldsymbol{x}}^{\mathrm{T}} - \bar{\boldsymbol{x}}\left(\frac{1}{n}\sum_{k=1}^{n}\boldsymbol{x}_k^{\mathrm{T}}\right) + \frac{1}{n}\sum_{k=1}^{n}\left(\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}}\right)$$

$$= \frac{1}{n}\sum_{k=1}^{n}\left(\boldsymbol{x}_k\boldsymbol{x}_k^{\mathrm{T}}\right) - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}} - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}} + \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}}$$

$$= \frac{1}{n}\left[\sum_{k=1}^{n}\left(\boldsymbol{x}_k\boldsymbol{x}_k^{\mathrm{T}}\right)\right] - \left(\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}}\right) = \frac{1}{n}\left(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\right)\begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \boldsymbol{x}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{pmatrix} - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}} = \frac{1}{n}\mathcal{X}^{\mathrm{T}}\mathcal{X} - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}}$$

描述性统计量 (Summary Statistics)

- 样本协方差矩阵：

$$\mathcal{S} = \frac{1}{n}\sum_{k=1}^{n}\left(\boldsymbol{x}_k - \bar{\boldsymbol{x}}\right)\left(\boldsymbol{x}_k - \bar{\boldsymbol{x}}\right)^{\mathrm{T}}$$

$$= \frac{1}{n}\mathcal{X}^{\mathrm{T}}\mathcal{X} - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}} \quad \Longleftarrow \quad \bar{\boldsymbol{x}} = \frac{1}{n}\mathcal{X}^{\mathrm{T}}\mathbf{1}_n$$

$$= \frac{1}{n}\mathcal{X}^{\mathrm{T}}\mathcal{X} - \frac{1}{n}\mathcal{X}^{\mathrm{T}}\mathbf{1}_n\left(\frac{1}{n}\mathcal{X}^{\mathrm{T}}\mathbf{1}_n\right)^{\mathrm{T}} = \frac{1}{n}\mathcal{X}^{\mathrm{T}}\mathcal{X} - \frac{1}{n^2}\mathcal{X}^{\mathrm{T}}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}\mathcal{X}$$

$$= \frac{1}{n}\left(\mathcal{X}^{\mathrm{T}}\mathcal{X} - \frac{1}{n}\mathcal{X}^{\mathrm{T}}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}\mathcal{X}\right)$$

Data matrix: $\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \boldsymbol{x}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{pmatrix}$

描述性统计量 (Summary Statistics)

- 样本协方差矩阵：

$$\mathcal{S} = \frac{1}{n}\sum_{k=1}^{n}\left(\boldsymbol{x}_k - \bar{\boldsymbol{x}}\right)\left(\boldsymbol{x}_k - \bar{\boldsymbol{x}}\right)^{\mathrm{T}}$$

$$= \frac{1}{n}\mathcal{X}^{\mathrm{T}}\mathcal{X} - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}}$$

$$= \frac{1}{n}\left(\mathcal{X}^{\mathrm{T}}\mathcal{X} - \frac{1}{n}\mathcal{X}^{\mathrm{T}}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}\mathcal{X}\right) = \frac{1}{n}\mathcal{X}^{\mathrm{T}}\left(\mathcal{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}\right)\mathcal{X}$$

$$= \frac{1}{n}\mathcal{X}^{\mathrm{T}}\mathcal{H}\mathcal{X}$$

Data matrix: $\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \boldsymbol{x}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{pmatrix}$

中心化矩阵 $\mathcal{H} = \mathcal{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$

描述性统计量 (Summary Statistics)

- 中心化矩阵 $\mathscr{H} = \mathscr{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$ 是对称幂等阵.

$$\mathscr{H}^{\mathrm{T}} = \left(\mathscr{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}\right)^{\mathrm{T}} = \mathscr{I}_n^{\mathrm{T}} - \frac{1}{n}\left(\mathbf{1}_n^{\mathrm{T}}\right)^{\mathrm{T}}\mathbf{1}_n^{\mathrm{T}} = \mathscr{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}} = \mathscr{H}$$

$$\mathscr{H}^2 = \left(\mathscr{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}\right)^2 = \left(\mathscr{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}\right)\left(\mathscr{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}\right)$$

$$= \mathscr{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}} + \left(\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}\right)\left(\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}\right)$$

$$= \mathscr{I}_n - \frac{2}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}} + \frac{1}{n^2}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$$

$\mathbf{1}_n^{\mathrm{T}}\mathbf{1}_n = n$

$$= \mathscr{I}_n - \frac{2}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}} + \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$$

$$= \mathscr{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}} = \mathscr{H}$$

描述性统计量 (Summary Statistics)

- 样本协方差矩阵 $\mathcal{S}$ 是非负定矩阵，即 $\mathcal{S} \geq 0$.

$\forall \boldsymbol{a} \in \mathbb{R}^p$,

$$\boldsymbol{a}^{\mathrm{T}} \mathcal{S} \boldsymbol{a} = \frac{1}{n} \boldsymbol{a}^{\mathrm{T}} \mathcal{X}^{\mathrm{T}} \mathcal{H} \mathcal{X} \boldsymbol{a} \quad \longleftarrow \quad \mathcal{S} = \frac{1}{n} \mathcal{X}^{\mathrm{T}} \mathcal{H} \mathcal{X}$$

$$= \frac{1}{n} \boldsymbol{a}^{\mathrm{T}} \mathcal{X}^{\mathrm{T}} \left(\mathcal{H}^{\mathrm{T}} \mathcal{H} \right) \mathcal{X} \boldsymbol{a} \quad \longleftarrow \quad \mathcal{H}\mathcal{H} = \mathcal{H}^{\mathrm{T}} \mathcal{H} = \mathcal{H}$$

$$= \frac{1}{n} \left(\boldsymbol{a}^{\mathrm{T}} \mathcal{X}^{\mathrm{T}} \mathcal{H}^{\mathrm{T}} \right) \left(\mathcal{H} \mathcal{X} \boldsymbol{a} \right)$$

$$= \frac{1}{n} \boldsymbol{y}^{\mathrm{T}} \boldsymbol{y} \quad \longleftarrow \quad \boldsymbol{y} = \mathcal{H} \mathcal{X} \boldsymbol{a}$$

$$= \frac{1}{n} \sum_{k=1}^{n} y_k^2 \geq 0$$

描述性统计量 (Summary Statistics)

- 由数理统计中的知识我们知道，一元情形中，$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ 是 $\sigma^2 = \mathrm{Var}(X)$ 的有偏估计量，而 $\frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ 是 $\sigma^2 = \mathrm{Var}(X)$ 的无偏估计量.
- 在多元情形中，可以证明 $\mathcal{S}_u = \frac{n}{n-1}\mathcal{S}$ 是总体协方差矩阵的一个无偏估计量.

描述性统计量 (Summary Statistics)

- 变量 i 与变量 j 之间的样本相关系数为

$$r_{X_iX_j} = \frac{s_{X_i,\,X_j}}{\sqrt{s_{X_i,\,X_i} \cdot s_{X_j,\,X_j}}}$$

- 相关矩阵 $\mathscr{R}$ 则为

$$\mathscr{R} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix} = \mathscr{D}^{-1/2}\mathcal{S}\,\mathscr{D}^{-1/2}$$

$$\mathscr{D}^{-1/2} = \operatorname{diag}\left(\sqrt{s_{X_1X_1}}\,,\sqrt{s_{X_2X_2}}\,,\,\cdots\,,\sqrt{s_{X_pX_p}}\right) = \begin{pmatrix} \sqrt{s_{X_1X_1}} & & & \\ & \sqrt{s_{X_2X_2}} & & \\ & & \ddots & \\ & & & \sqrt{s_{X_pX_p}} \end{pmatrix}$$

描述性统计量 (Summary Statistics)

- 例: 套头衫数据集的样本协方差矩阵的计算.

```
pullover
pullover.mean <- sapply(pullover, mean)
pullover.mean
```

```
> pullover.mean
   X1    X2    X3    X4
172.7 104.6 104.0  93.8
```

$$\Longrightarrow \quad \bar{\boldsymbol{x}} = \begin{pmatrix} 172.7 \\ 104.6 \\ 104.0 \\ 93.8 \end{pmatrix}$$

$\mathcal{X} =$

```
> pullover
    X1  X2  X3  X4
1  230 125 200 109
2  181  99  55 107
3  165  97 105  98
4  150 115  85  71
5   97 120   0  82
6  192 100 150 103
7  181  80  85 111
8  189  90 120  93
9  172  95 110  86
10 170 125 130  78
```

```
pullover <- as.matrix(pullover)
pullover.mean <- as.matrix(pullover.mean)
pullover.cov <- (1/10) * t(pullover) %*% pullover - pullover.mean %*% t(pullover.mean)
pullover.cov
```

$$\mathcal{S} = \begin{pmatrix} 1037.21 & -80.02 & 1430.7 & 271.44 \\ -80.02 & 219.84 & 92.1 & -91.58 \\ 1430.70 & 92.10 & 2624.0 & 210.30 \\ 271.44 & -91.58 & 210.3 & 177.36 \end{pmatrix}$$

$$\mathcal{S} = \frac{1}{n}\mathcal{X}^{\mathrm{T}}\mathcal{X} - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}}$$

描述性统计量 (Summary Statistics)

- 例: 套头衫数据集的样本协方差矩阵的计算.
  - 协方差矩阵的无偏估计为

```
pullover.cov.u <- (10/9) * pullover.cov
pullover.cov.u
```

```
> pullover.cov.u
            X1          X2          X3          X4
X1 1152.45556  -88.91111 1589.6667  301.6000
X2  -88.91111  244.26667  102.3333 -101.7556
X3 1589.66667  102.33333 2915.5556  233.6667
X4  301.60000 -101.75556  233.6667  197.0667
```

$$\Longrightarrow \quad \mathcal{S}_u = \frac{n}{n-1}\mathcal{S} = \begin{pmatrix} 1152.45556 & -88.91111 & 1589.6667 & 301.6000 \\ -88.91111 & 244.26667 & 102.3333 & -101.7556 \\ 1589.66667 & 102.33333 & 2915.5556 & 233.6667 \\ 301.60000 & -101.75556 & 233.6667 & 197.0667 \end{pmatrix}$$

```
pullover.cov.u2 <- cov(pullover)
pullover.cov.u2
```

```
> pullover.cov.u2
            X1          X2          X3          X4
X1 1152.45556  -88.91111 1589.6667  301.6000
X2  -88.91111  244.26667  102.3333 -101.7556
X3 1589.66667  102.33333 2915.5556  233.6667
X4  301.60000 -101.75556  233.6667  197.0667
```

描述性统计量 (Summary Statistics)

- 例: 套头衫数据集的样本协方差矩阵的计算.
  - 样本相关矩阵为

```
pullover.cor <- cor(pullover)
round(pullover.cor, digits = 2)
```

```
> round(pullover.cor, digits = 2)
      X1    X2   X3    X4
X1  1.00 -0.17 0.87  0.63
X2 -0.17  1.00 0.12 -0.46
X3  0.87  0.12 1.00  0.31
X4  0.63 -0.46 0.31  1.00
```

$$\Longrightarrow \quad \mathscr{R} = \begin{pmatrix} 1.00 & -0.17 & 0.87 & 0.63 \\ -0.17 & 1.00 & 0.12 & -0.46 \\ 0.87 & 0.12 & 1.00 & 0.31 \\ 0.63 & -0.46 & 0.31 & 1.00 \end{pmatrix}$$

描述性统计量 (Summary Statistics)

- 线性变换：$\boldsymbol{Y} = \mathscr{A}_{q\times p}\boldsymbol{X}$

随机向量：$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$

数据矩阵：$\mathscr{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \triangleq \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \boldsymbol{x}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{pmatrix}$

变换矩阵：$\mathscr{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{q1} & a_{q2} & \cdots & a_{qp} \end{pmatrix}$

变换后得随机向量：$\boldsymbol{Y} = \mathscr{A}\boldsymbol{X} = \begin{pmatrix} a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ \vdots \\ a_{q1}X_1 + a_{q2}X_2 + \cdots + a_{qp}X_p \end{pmatrix} \triangleq \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_q \end{pmatrix}$

变换后的数据矩阵：$\mathscr{Y} = \mathscr{X}\mathscr{A}^{\mathrm{T}} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1q} \\ y_{21} & y_{22} & \cdots & y_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nq} \end{pmatrix} \triangleq \begin{pmatrix} \boldsymbol{y}_1^{\mathrm{T}} \\ \boldsymbol{y}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{y}_n^{\mathrm{T}} \end{pmatrix}$

描述性统计量 (Summary Statistics)

- 线性变换：$\boldsymbol{Y} = \mathscr{A}_{q\times p}\boldsymbol{X}$

变换后的数据矩阵: $\mathscr{Y} = \mathscr{X}\mathscr{A}^{\mathrm{T}}$

$$\bar{\boldsymbol{y}} = \frac{1}{n}\mathscr{Y}^{\mathrm{T}}\mathbf{1}_n = \frac{1}{n}\left(\mathscr{X}\mathscr{A}^{\mathrm{T}}\right)^{\mathrm{T}}\mathbf{1}_n = \frac{1}{n}\mathscr{A}\mathscr{X}^{\mathrm{T}}\mathbf{1}_n$$

$$= \mathscr{A}\left(\frac{1}{n}\mathscr{X}^{\mathrm{T}}\mathbf{1}_n\right) = \mathscr{A}\,\bar{\boldsymbol{x}}$$

$$\mathscr{S}_{\mathscr{Y}} = \frac{1}{n}\mathscr{Y}^{\mathrm{T}}\mathscr{H}\mathscr{Y} = \frac{1}{n}\left(\mathscr{X}\mathscr{A}^{\mathrm{T}}\right)^{\mathrm{T}}\mathscr{H}\left(\mathscr{X}\mathscr{A}^{\mathrm{T}}\right)$$

中心化矩阵 $\mathscr{H} = \mathscr{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$

$$= \frac{1}{n}\mathscr{A}\mathscr{X}^{\mathrm{T}}\mathscr{H}\mathscr{X}\mathscr{A}^{\mathrm{T}}$$

$$= \mathscr{A}\left(\frac{1}{n}\mathscr{X}^{\mathrm{T}}\mathscr{H}\mathscr{X}\right)\mathscr{A}^{\mathrm{T}}$$

$$= \mathscr{A}\mathscr{S}_{\mathscr{X}}\mathscr{A}^{\mathrm{T}}$$

描述性统计量 (Summary Statistics)

- 线性变换：$\boldsymbol{Y}=\mathscr{A}_{q\times p}\boldsymbol{X}$ $\Longrightarrow$ $\begin{cases}\bar{\boldsymbol{y}}=\mathscr{A}\bar{\boldsymbol{x}}\\ \mathcal{S}_{\mathcal{Y}}=\mathscr{A}\mathcal{S}_{\mathcal{X}}\mathscr{A}^{\mathrm{T}}\end{cases}$

  - 如果是非齐次的线性变换 $\boldsymbol{Y}=\mathscr{A}\boldsymbol{X}+\boldsymbol{b}$，$\boldsymbol{b}\in\mathbb{R}^q$

$$\Longrightarrow \begin{cases}\bar{\boldsymbol{y}}=\mathscr{A}\bar{\boldsymbol{x}}+\boldsymbol{b}\\ \mathcal{S}_{\mathcal{Y}}=\mathscr{A}\mathcal{S}_{\mathcal{X}}\mathscr{A}^{\mathrm{T}}\end{cases}$$

  - 特别：$q=1$ 时，$Y=\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}$，其中 $\boldsymbol{a}\in\mathbb{R}^p$，变换后的数据 $\boldsymbol{y}_{n\times 1}=\mathscr{X}\boldsymbol{a}$，

即 $y_i=\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_i, i=1,2,\ldots,n$

$$\Longrightarrow \begin{cases}\bar{y}=\boldsymbol{a}^{\mathrm{T}}\bar{\boldsymbol{x}}\\ \mathcal{S}_{y}=\boldsymbol{a}^{\mathrm{T}}\mathcal{S}_{\mathcal{X}}\boldsymbol{a}\end{cases}$$

描述性统计量 (Summary Statistics)

- 例: 设 $\mathcal{X}$ 是套头衫的数据集. 经理欲计算广告费用 (X_3) 加促销员费用 (X_4) 的平均支出. 假设促销员的时薪为 10 欧元.

```
pullover
```

```
> pullover
      X1  X2  X3  X4
 [1,] 230 125 200 109
 [2,] 181  99  55 107
 [3,] 165  97 105  98
 [4,] 150 115  85  71
 [5,]  97 120   0  82
 [6,] 192 100 150 103
 [7,] 181  80  85 111
 [8,] 189  90 120  93
 [9,] 172  95 110  86
[10,] 170 125 130  78
```

$$Y = X_3 + 10X_4 = (0 \quad 0 \quad 1 \quad 10)\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} = \mathcal{A}X$$

```
A <- matrix(c(0, 0, 1, 10), nrow=1)
A %*% pullover.mean
```

```
> A %*% pullover.mean
     [,1]
[1,] 1042
```

$$\Longrightarrow \quad \bar{y} = \mathcal{A}\bar{x} = (0 \quad 0 \quad 1 \quad 10)\begin{pmatrix} 172.7 \\ 104.6 \\ 104.0 \\ 93.8 \end{pmatrix} = 1042.0$$

描述性统计量 (Summary Statistics)

- 例: 设 $\mathcal{X}$ 是套头衫的数据集. 经理欲计算广告费用 (X_3) 加促销员费用 (X_4) 的平均支出. 假设促销员的时薪为 10 欧元.

```
pullover
```

```
> pullover
      X1  X2  X3  X4
 [1,] 230 125 200 109
 [2,] 181  99  55 107
 [3,] 165  97 105  98
 [4,] 150 115  85  71
 [5,]  97 120   0  82
 [6,] 192 100 150 103
 [7,] 181  80  85 111
 [8,] 189  90 120  93
 [9,] 172  95 110  86
[10,] 170 125 130  78
```

$$Y = X_3 + 10X_4 = (0 \quad 0 \quad 1 \quad 10)\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} = \mathcal{A}X$$

```
S_Y <- A %*% pullover.cov.u2 %*% t(A)
S_Y
```

```
> S_Y
          [,1]
[1,] 27295.56
```

$$\Longrightarrow \quad \mathcal{S}_y = \mathcal{A}\mathcal{S}_x\mathcal{A}^{\mathrm{T}}$$

$$= (0 \quad 0 \quad 1 \quad 10)\begin{pmatrix} 1152.45556 & -88.91111 & 1589.6667 & 301.6000 \\ -88.91111 & 244.26667 & 102.3333 & -101.7556 \\ 1589.66667 & 102.33333 & 2915.5556 & 233.6667 \\ 301.60000 & -101.75556 & 233.6667 & 197.0667 \end{pmatrix}\begin{pmatrix} 0 \\ 0 \\ 1 \\ 10 \end{pmatrix}$$

$$= 27295.56$$

描述性统计量 (Summary Statistics)

- Mahalanobis 变换
  - 线性变换的一种特殊情形

$$z_i = \mathcal{S}^{-1/2}\left(x_i - \bar{x}\right), \quad i = 1, 2, \ldots, n$$

$$\mathcal{Z} = \left(z_1, z_2, \ldots, z_n\right)^{\mathrm{T}} = \begin{pmatrix} z_1^{\mathrm{T}} \\ z_2^{\mathrm{T}} \\ \vdots \\ z_n^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} \left(x_1 - \bar{x}\right)^{\mathrm{T}} \mathcal{S}^{-1/2} \\ \left(x_2 - \bar{x}\right)^{\mathrm{T}} \mathcal{S}^{-1/2} \\ \vdots \\ \left(x_n - \bar{x}\right)^{\mathrm{T}} \mathcal{S}^{-1/2} \end{pmatrix}$$

$$= \begin{pmatrix} x_1^{\mathrm{T}} - \bar{x}^{\mathrm{T}} \\ x_2^{\mathrm{T}} - \bar{x}^{\mathrm{T}} \\ \vdots \\ x_n^{\mathrm{T}} - \bar{x}^{\mathrm{T}} \end{pmatrix} \mathcal{S}^{-1/2} = \begin{pmatrix} x_1^{\mathrm{T}} \\ x_2^{\mathrm{T}} \\ \vdots \\ x_n^{\mathrm{T}} \end{pmatrix} \mathcal{S}^{-1/2} - \begin{pmatrix} \bar{x}^{\mathrm{T}} \\ \bar{x}^{\mathrm{T}} \\ \vdots \\ \bar{x}^{\mathrm{T}} \end{pmatrix} \mathcal{S}^{-1/2}$$

$$= \mathcal{X}\mathcal{S}^{-1/2} - \mathbf{1}_n \bar{x}^{\mathrm{T}} \mathcal{S}^{-1/2} = \mathcal{X}\mathcal{S}^{-1/2} - \mathbf{1}_n \left(\frac{1}{n}\mathcal{X}^{\mathrm{T}}\mathbf{1}_n\right)^{\mathrm{T}} \mathcal{S}^{-1/2}$$

$$= \mathcal{X}\mathcal{S}^{-1/2} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}\mathcal{X}\mathcal{S}^{-1/2} = \mathcal{H}\mathcal{X}\mathcal{S}^{-1/2}$$

$$\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \triangleq \begin{pmatrix} x_1^{\mathrm{T}} \\ x_2^{\mathrm{T}} \\ \vdots \\ x_n^{\mathrm{T}} \end{pmatrix}$$

$$\bar{x} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix} = \frac{1}{n}\mathcal{X}^{\mathrm{T}}\mathbf{1}_n$$

$$\mathcal{S} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(x_i - \bar{x}\right)^{\mathrm{T}}$$

$$= \frac{1}{n}\mathcal{X}^{\mathrm{T}}\mathcal{X} - \bar{x}\bar{x}^{\mathrm{T}}$$

$$= \frac{1}{n}\mathcal{X}^{\mathrm{T}}\mathcal{H}\mathcal{X}$$

$$\mathcal{S} \geq 0$$

$$\mathcal{H} = \mathcal{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$$

描述性统计量 (Summary Statistics)

- Mahalanobis 变换
  - 线性变换的一种特殊情形

$$z_i = \mathcal{S}^{-1/2}\left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}\right), \quad i = 1, 2, \ldots, n$$

$$\mathcal{Z} = \left(\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n\right)^{\mathrm{T}} = \mathcal{H}\mathcal{X}\mathcal{S}^{-1/2}$$

$$\mathcal{S}_{\mathcal{Z}} = \frac{1}{n}\mathcal{Z}^{\mathrm{T}}\mathcal{H}\mathcal{Z} = \frac{1}{n}\left(\mathcal{H}\mathcal{X}\mathcal{S}^{-1/2}\right)^{\mathrm{T}}\mathcal{H}\left(\mathcal{H}\mathcal{X}\mathcal{S}^{-1/2}\right)$$

$$= \frac{1}{n}\mathcal{S}^{-1/2}\mathcal{X}^{\mathrm{T}}\mathcal{H}^{\mathrm{T}}\mathcal{H}\mathcal{H}\mathcal{X}\mathcal{S}^{-1/2}$$

对称、幂等

$$= \frac{1}{n}\mathcal{S}^{-1/2}\mathcal{X}^{\mathrm{T}}\mathcal{H}\mathcal{X}\mathcal{S}^{-1/2}$$

$$= \mathcal{S}^{-1/2}\mathcal{S}\mathcal{S}^{-1/2} = \mathcal{I}_p$$

$$\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \triangleq \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \boldsymbol{x}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{pmatrix}$$

$$\bar{\boldsymbol{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix} = \frac{1}{n}\mathcal{X}^{\mathrm{T}}\mathbf{1}_n$$

$$\mathcal{S} = \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}\right)\left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}\right)^{\mathrm{T}}$$

$$= \frac{1}{n}\mathcal{X}^{\mathrm{T}}\mathcal{X} - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}}$$

$$= \frac{1}{n}\mathcal{X}^{\mathrm{T}}\mathcal{H}\mathcal{X}$$

$$\mathcal{H} = \mathcal{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$$

描述性统计量 (Summary Statistics)

- Mahalanobis 变换
  - 线性变换的一种特殊情形

$$z_i = \mathcal{S}^{-1/2}\left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}\right), \quad i = 1, 2, \ldots, n$$

$$\mathcal{Z} = \left(z_1, z_2, \ldots, z_n\right)^{\mathrm{T}} = \mathcal{H}\mathcal{X}\mathcal{S}^{-1/2}$$

$$\mathcal{S}_{\mathcal{Z}} = \mathcal{I}_p$$

  - 因此，Mahalanobis 变换消除了变量之间的相关性，并对每一个变量实施了标准化变换.

两个变量的线性模型

- 简单线性回归模型
  - 两个变量 X 与 Y 满足

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

$$\mathrm{E}(\varepsilon_i) = 0, \ \mathrm{Var}(\varepsilon_i) = \sigma^2, \text{ or } \varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

  - 估计

$$\left(\hat{\beta}_0, \hat{\beta}_1\right) = \arg\min_{(\beta_0, \beta_1)} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2 \Longrightarrow \hat{\beta}_1 = \frac{s_{XY}}{s_{XX}}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\widehat{\sigma^2} = \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - \widehat{y_i}\right)^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

$$\mathrm{Var}\left(\hat{\beta}_1\right) = \frac{\widehat{\sigma^2}}{n \cdot s_{XX}} \Longrightarrow \mathrm{SE}\left(\hat{\beta}_1\right) = \sqrt{\mathrm{Var}\left(\hat{\beta}_1\right)} = \frac{\hat{\sigma}}{\sqrt{n \cdot s_{XX}}}$$

两个变量的线性模型

- 简单线性回归模型
  - 假设检验 $\quad \mathrm{H}_0: \beta_1 = 0 \leftrightarrow \mathrm{H}_1: \beta_1 \neq 0$

$$t = \frac{\hat{\beta}_1}{\mathrm{SE}\left(\hat{\beta}_1\right)} \overset{\mathrm{H}_0}{\sim} t_{n-2} \Longrightarrow \text{当 } |t| \geq t_{n-2}\left(\frac{\alpha}{2}\right) \text{ 时拒绝 } \mathrm{H}_0$$

$$\Longrightarrow \text{当 } p\text{值} = 2 \times P\left(t_{n-2} > |t|\right) < \alpha \text{ 时拒绝 } \mathrm{H}_0$$

  - 可以证明

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 + \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

$$\text{总平方和} = \text{回归平方和} + \text{残差平方和}$$

  - 决定系数

$$r^2 = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} = \frac{\text{回归平方和}}{\text{总平方和}} = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

两个变量的线性模型

- 例: 我们对套头衫数据集应用线性回归模型.

```
pullover
plot(X1 ~ X2, data = pullover, xlab = "Price (X2)", ylab = "Sales (X1)", pch = 16, cex = 1.5)
pullover.lm <- lm(X1 ~ X2, data = pullover)
abline(pullover.lm, col = "red", lwd = 2)
summary(pullover.lm)
```

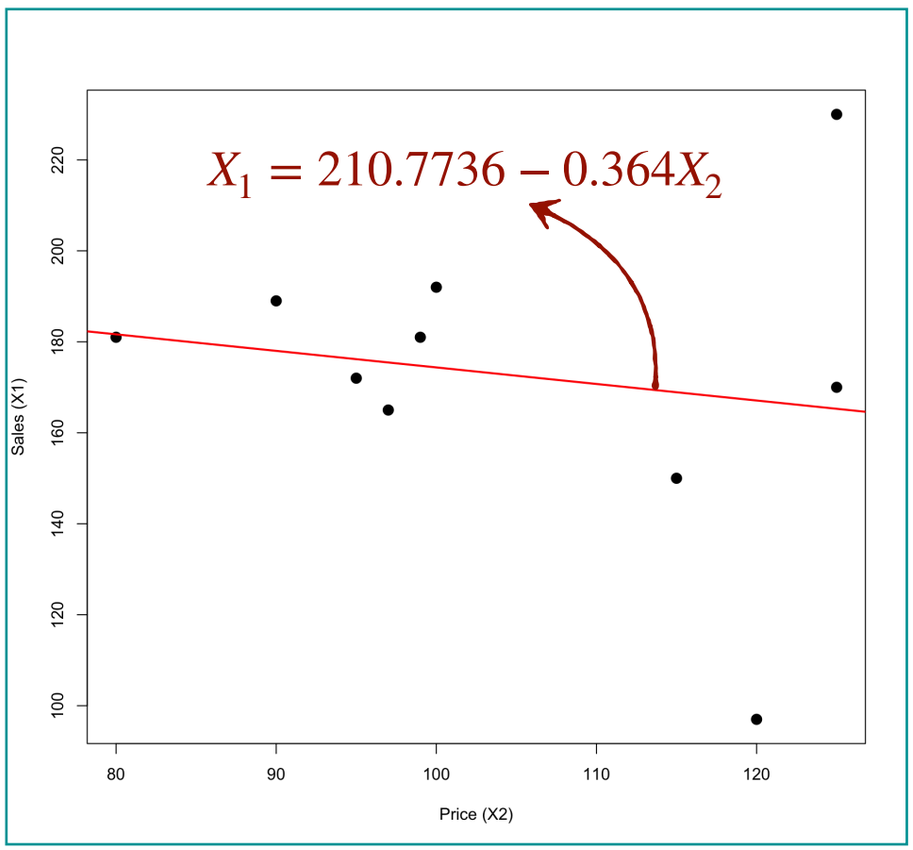
```
> summary(pullover.lm)

Call:
lm(formula = X1 ~ X2, data = pullover)

Residuals:
    Min      1Q  Median      3Q     Max
-70.095  -8.898   2.036   9.805  64.725

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 210.7736    79.9837   2.635   0.0299 *
X2           -0.3640     0.7571  -0.481   0.6435
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35.5 on 8 degrees of freedom
Multiple R-squared:  0.02808,   Adjusted R-squared:  -0.09341
F-statistic: 0.2311 on 1 and 8 DF,  p-value: 0.6435
```

$H_0: \beta_1 = 0 \leftrightarrow H_1: \beta_1 \neq 0$

接受 H_0

$r^2 = 0.02808$

```
> pullover
    X1  X2  X3  X4
1  230 125 200 109
2  181  99  55 107
3  165  97 105  98
4  150 115  85  71
5   97 120   0  82
6  192 100 150 103
7  181  80  85 111
8  189  90 120  93
9  172  95 110  86
10 170 125 130  78
```



单因子方差分析 (Simple Analysis of Variance)

- 单因子方差分析 (ANOVA)
  - 假设响应变量 y 的平均值仅受单一因子影响.
  - 该因子有 p 个不同值 (水平).
  - 对该因子的每一个水平观测 m 次.
  - 假设所有的观测结果相互独立.

$$\Leftrightarrow \begin{cases} \mathrm{H}_0 : & \alpha_1 = \alpha_2 = \cdots = \alpha_p = 0 \\ \mathrm{H}_1 : & \text{at least one of } \alpha_j \neq 0 \end{cases}$$

  - 目的：分析不同水平的均值是否有显著差异

$$\begin{cases} \mathrm{H}_0 : & \mu_1 = \mu_2 = \cdots = \mu_p = \mu \\ \mathrm{H}_1 : & \mu_k \neq \mu_l \ \text{ for some } k \text{ and } l \end{cases}$$

$$y_{ij} = \mu_j + \varepsilon_{ij} = \left(\mu + \alpha_j\right) + \varepsilon_{ij}, \quad i = 1, 2, \ldots, m; \ j = 1, 2, \ldots, p$$

Sample element i	Factor levels j				
	1	⋯	j	⋯	p
1	y_{11}	⋯	y_{1j}	⋯	y_{1p}
2	y_{21}	⋯	y_{2j}	⋯	y_{2p}
⋮	⋮		⋮		⋮
i	y_{i1}	⋯	y_{ij}	⋯	y_{ip}
⋮	⋮		⋮		⋮
m	y_{m1}	⋯	y_{mj}	⋯	y_{mp}

单因子方差分析 (Simple Analysis of Variance)

- 单因子方差分析 (ANOVA)

$$y_{ij} = \mu_j + \varepsilon_{ij} = (\mu + \alpha_j) + \varepsilon_{ij}, \quad i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, p$$

第 j 个水平的均值

$$\varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

  - 当 $m = 1, p = n, \mu_j = \beta_0 + \beta_1 x_j$ 时，上述模型就是线性回归模型，其中 x_j 是因子的第 j 个水平的值.

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \ldots, n$$

Sample element i	Factor levels j				
	x_1	$\cdots$	x_j	$\cdots$	x_n
1	y_1	$\cdots$	y_j	$\cdots$	y_n

单因子方差分析 (Simple Analysis of Variance)

- 例: 销售套头衫的公司欲分析三种营销策略的效果.

  1. 本地报纸作广告
  2. 有促销员
  3. 橱窗里的奢华展示

将这些策略在 10 家不同的商店进行了试验.

```
Sales <- c(9, 11, 10, 12, 7, 11, 12, 10, 11, 13, 10, 15, 11, 15, 15, 13, 7, 15, 113, 10, 18,
           14, 17, 9, 14, 17, 16, 14, 17, 15)
Strategy <- as.factor(c(rep(1, 10), rep(2, 10), rep(3, 10)))
Pullover <- data.frame(Sales = Sales, Strategy = Strategy)
Pullover
```

- 公司欲分析这三种营销策略的平均效果是否相同，或是否存在差异.

$$\begin{cases} \mathrm{H}_0 : & \mu_1 = \mu_2 = \mu_3 = \mu \\ \mathrm{H}_1 : & \mu_k \neq \mu_l \quad \text{for some } k \text{ and } l \end{cases}$$

$$\Longleftrightarrow \quad \begin{cases} \mathrm{H}_0 : & \alpha_1 = \alpha_2 = \alpha_3 = 0 \\ \mathrm{H}_1 : & \text{at least one of } \alpha_j \neq 0 \end{cases}$$

```
> Pullover
   Sales Strategy
1      9        1
2     11        1
3     10        1
4     12        1
5      7        1
6     11        1
7     12        1
8     10        1
9     11        1
10    13        1
11    10        2
12    15        2
13    11        2
14    15        2
15    15        2
16    13        2
17     7        2
18    15        2
19    13        2
20    10        2
21    18        3
22    14        3
23    17        3
24     9        3
25    14        3
26    17        3
27    16        3
28    14        3
29    17        3
30    15        3
```

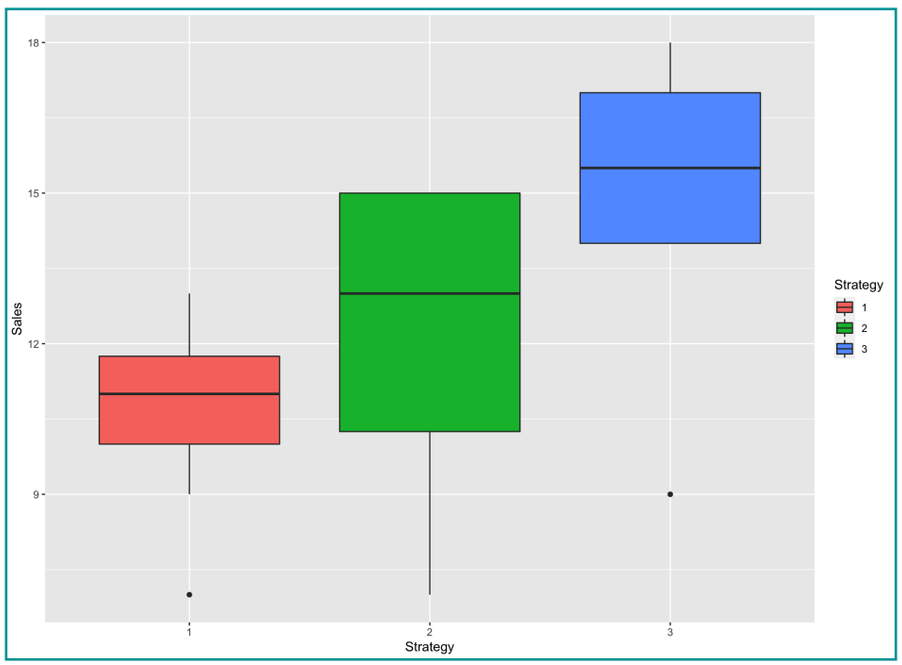
单因子方差分析 (Simple Analysis of Variance)

- 例: 销售套头衫的公司欲分析三种营销策略的效果.

  1. 本地报纸作广告
  2. 有促销员
  3. 橱窗里的奢华展示

将这些策略在 10 家不同的商店进行了试验.

```
## 可视化
library(ggplot2)
p1 <- ggplot(data = Pullover) +
  geom_boxplot(aes(x = Strategy, y = Sales, fill = Strategy))
p1
```




```
> Pullover
   Sales Strategy
1      9        1
2     11        1
3     10        1
4     12        1
5      7        1
6     11        1
7     12        1
8     10        1
9     11        1
10    13        1
11    10        2
12    15        2
13    11        2
14    15        2
15    15        2
16    13        2
17     7        2
18    15        2
19    13        2
20    10        2
21    18        3
22    14        3
23    17        3
24     9        3
25    14        3
26    17        3
27    16        3
28    14        3
29    17        3
30    15        3
```

单因子方差分析 (Simple Analysis of Variance)

- 例: 销售套头衫的公司欲分析三种营销策略的效果.

  1. 本地报纸作广告
  2. 有促销员
  3. 橱窗里的奢华展示

将这些策略在 10 家不同的商店进行了试验.

```
## 因子 (营销策略) 各水平的均值、标准差
library(dplyr)
group_by(Pullover, Strategy) %>%
  summarise(
    n_i = n(),
    hat_mean_i = mean(Sales),
    hat_sd_i = sd(Sales)
  )
```

```
# A tibble: 3 × 4
  Strategy   n_i hat_mean_i hat_sd_i
  <fct>    <int>      <dbl>    <dbl>
1 1           10       10.6     1.71
2 2           10       12.4     2.80
3 3           10       15.1     2.60
```

```
> Pullover
   Sales Strategy
1      9        1
2     11        1
3     10        1
4     12        1
5      7        1
6     11        1
7     12        1
8     10        1
9     11        1
10    13        1
11    10        2
12    15        2
13    11        2
14    15        2
15    15        2
16    13        2
17     7        2
18    15        2
19    13        2
20    10        2
21    18        3
22    14        3
23    17        3
24     9        3
25    14        3
26    17        3
27    16        3
28    14        3
29    17        3
30    15        3
```

单因子方差分析 (Simple Analysis of Variance)

- 例: 销售套头衫的公司欲分析三种营销策略的效果.

$$\begin{cases} \mathrm{H}_0 : & \mu_1 = \mu_2 = \mu_3 = \mu \\ \mathrm{H}_1 : & \mu_k \neq \mu_l \quad \text{for some } k \text{ and } l \end{cases} \iff \begin{cases} \mathrm{H}_0 : & \alpha_1 = \alpha_2 = \alpha_3 = 0 \\ \mathrm{H}_1 : & \text{at least one of } \alpha_j \neq 0 \end{cases}$$

  - 模型拟合

```
sales.lm <- lm(Sales ~ Strategy, data = Pullover)
summary(sales.lm)
```

```
> summary(sales.lm)

Call:
lm(formula = Sales ~ Strategy, data = Pullover)

Residuals:
   Min     1Q Median     3Q    Max
  -6.1   -1.1    0.4    1.9    2.9

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.6000     0.7642  13.870 8.44e-14 ***
Strategy2     1.8000     1.0808   1.665 0.107392
Strategy3     4.5000     1.0808   4.164 0.000287 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.417 on 27 degrees of freedom
Multiple R-squared:  0.3942,    Adjusted R-squared:  0.3493
F-statistic: 8.783 on 2 and 27 DF,  p-value: 0.001153
```

$\hat{\mu}_1$

$\hat{\mu}_2 - \hat{\mu}_1$

$\hat{\mu}_3 - \hat{\mu}_1$

单因子方差分析 (Simple Analysis of Variance)

Moving to Higher Dimensions

Sample element i	Factor levels j				
	1	⋯	j	⋯	p
1	y_{11}	⋯	y_{1j}	⋯	y_{1p}
2	y_{21}	⋯	y_{2j}	⋯	y_{2p}
⋮	⋮		⋮		⋮
i	y_{i1}	⋯	y_{ij}	⋯	y_{ip}
⋮	⋮		⋮		⋮
m	y_{m1}	⋯	y_{mj}	⋯	y_{mp}

- 单因子方差分析 (ANOVA)
  - 总偏差平方和 (H_0 为真时)

$$SS_{reduced} = \sum_{i=1}^{m}\sum_{j=1}^{p}\left(y_{ij} - \bar{y}\right)^2$$

$$\bar{y} = \frac{1}{mp}\sum_{i=1}^{m}\sum_{j=1}^{p} y_{ij}$$ 总均值

  - H_1 为真时的偏差平方和

$$SS_{full} = \sum_{i=1}^{m}\sum_{j=1}^{p}\left(y_{ij} - \bar{y}_j\right)^2$$

$$\bar{y}_j = \frac{1}{m}\sum_{i=1}^{m} y_{ij}$$ 第 j 个水平的均值

  - 可以证明

$$\sum_{i=1}^{m}\sum_{j=1}^{p}\left(y_{ij} - \bar{y}\right)^2 = m\sum_{j=1}^{p}\left(\bar{y}_j - \bar{y}\right)^2 + \sum_{i=1}^{m}\sum_{j=1}^{p}\left(y_{ij} - \bar{y}_j\right)^2$$

$SS_{treatment}$

总偏差平方和 = 组间偏差平方和 + 组内偏差平方和

$$SS_{reduced} = SS_{treatment} + SS_{full}$$

单因子方差分析 (Simple Analysis of Variance)

- 单因子方差分析 (ANOVA)
  - 单因子方差分析模型中的大部分信息都包含在方差分析表中

Source	SS	df	MS	E(MS)
Treatment	$\mathrm{SS}_{\text{treatment}} = m\sum_{j=1}^{p}\left(\bar{y}_j - \bar{y}\right)^2$	$p-1$	$\dfrac{\mathrm{SS}_{\text{treatment}}}{p-1}$	$\sigma^2 + \dfrac{m}{p-1}\sum_{j=1}^{p}\alpha_j^2$
Error	$\mathrm{SS}_{\text{full}} = \sum_{i=1}^{m}\sum_{j=1}^{p}\left(y_{ij} - \bar{y}_j\right)^2$	$p(m-1)$	$\dfrac{\mathrm{SS}_{\text{full}}}{p(m-1)}$	σ^2

  - 方差分析表中的行通常是独立的.
  - 当 H_0： $\mu_1 = \mu_2 = \cdots = \mu_p = \mu$ 或 H_0： $\alpha_1 = \alpha_2 = \cdots = \alpha_p = 0$ 为真时

$$F = \frac{\mathrm{MS}_{\text{treatment}}}{\mathrm{MS}_{\text{full}}} = \frac{\dfrac{\mathrm{SS}_{\text{treatment}}}{df_{\text{treatment}}}}{\dfrac{\mathrm{SS}_{\text{full}}}{df_{\text{full}}}} \sim F_{df_{\text{treatment}},\, df_{\text{full}}}$$

  - 显著水平取 α，则当 $F \geq F_{df_{\text{treatment}},\, df_{\text{full}}}(\alpha)$ 时拒绝 H_0 .

单因子方差分析 (Simple Analysis of Variance)

- 例: 销售套头衫的公司欲分析三种营销策略的效果.

$$\begin{cases} \mathrm{H}_0 : & \mu_1 = \mu_2 = \mu_3 = \mu \\ \mathrm{H}_1 : & \mu_k \neq \mu_l \quad \text{for some } k \text{ and } l \end{cases} \iff \begin{cases} \mathrm{H}_0 : & \alpha_1 = \alpha_2 = \alpha_3 = 0 \\ \mathrm{H}_1 : & \text{at least one of } \alpha_j \neq 0 \end{cases}$$

```
anova(sales.lm)
```

```
> anova(sales.lm)
Analysis of Variance Table

Response: Sales
           Df Sum Sq Mean Sq F value   Pr(>F)
Strategy    2  102.6  51.300  8.7831 0.001153 **
Residuals 27  157.7   5.841
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- 我们可以用 $\mathfrak{R}$ 中的 aov() 函数来作方差分析

```
sales.aov <- aov(Sales ~ Strategy, data = Pullover)
summary(sales.aov)
```

```
> summary(sales.aov)
             Df Sum Sq Mean Sq F value  Pr(>F)
Strategy      2  102.6   51.30   8.783 0.00115 **
Residuals    27  157.7    5.84
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

多重线性模型 (Multiple Linear Model)

- 响应变量 Y，数据向量 $\boldsymbol{y}=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ 是响应变量 Y 的一组观测值.

- 解释变量 $\boldsymbol{X}=\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$，数据矩阵 $\mathcal{X}=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \\ \uparrow & \uparrow & \cdots & \uparrow \\ X_1 & X_2 & \cdots & X_p \end{pmatrix}$ 是解释变量 $\boldsymbol{X}$ 的一组观测值.

多重线性模型 (Multiple Linear Model)

- 线性模型 (无截距项) 可以表示为

$$\boldsymbol{y} = \mathcal{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

误差项： $\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$

回归系数： $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$

- 已有理论的重要应用：最小二乘拟合.

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \mathcal{X}\boldsymbol{\beta})^{\mathrm{T}} (\boldsymbol{y} - \mathcal{X}\boldsymbol{\beta})$$

  - 如果矩阵 $(\mathcal{X}^{\mathrm{T}}\mathcal{X})$ 满秩，因此可逆.

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

  - 拟合值

$$\hat{\boldsymbol{y}} = \mathcal{X}\hat{\boldsymbol{\beta}} = \mathcal{X}(\mathcal{X}^{\mathrm{T}}\mathcal{X})^{-1}\mathcal{X}^{\mathrm{T}}\boldsymbol{y} = \mathcal{P}\,\boldsymbol{y}$$

向量 $\boldsymbol{y}$ 在列空间 $C(\mathcal{X})$ 上的投影

多重线性模型 (Multiple Linear Model)

- 线性模型 (无截距项) 可以表示为

$$\boldsymbol{y} = \mathcal{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

误差项： $\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$

回归系数： $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$

- 已有理论的重要应用：最小二乘拟合.

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \mathcal{X}\boldsymbol{\beta})^{\mathrm{T}} (\boldsymbol{y} - \mathcal{X}\boldsymbol{\beta})$$

  - 如果矩阵 $(\mathcal{X}^{\mathrm{T}}\mathcal{X})$ 满秩，因此可逆.

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

  - 最小二乘残差为

向量 $\boldsymbol{y}$ 在 $C(\mathcal{X})$ 的正交补上的投影

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \mathcal{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \mathcal{P}\boldsymbol{y} = (\mathcal{I}_n - \mathcal{P})\boldsymbol{y} = \mathcal{Q}\boldsymbol{y}$$

多重线性模型 (Multiple Linear Model)

注 3.5 具有截距项 β_0 的线性模型也有类似的形式. 模型的表达式为

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i\,, \quad i = 1, 2, \ldots, n$$

亦可表示为矩阵形式

$$\boldsymbol{y} = \mathscr{X}^*\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$$

其中 $\mathscr{X}^* = \left(\mathbf{1}_n \;\; \mathscr{X}\right)$ (给数据矩阵添加了一列1向量). 从而我们有

$$\widehat{\boldsymbol{\beta}^*} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}} \end{pmatrix} = \left(\mathscr{X}^{*\mathrm{T}}\mathscr{X}^*\right)^{-1}\mathscr{X}^{*\mathrm{T}}\boldsymbol{y}$$

多重线性模型 (Multiple Linear Model)

- 例：我们依然讨论套头衫数据集的例子.

```
pullover
```

```
> pullover
    X1  X2  X3  X4
1  230 125 200 109
2  181  99  55 107
3  165  97 105  98
4  150 115  85  71
5   97 120   0  82
6  192 100 150 103
7  181  80  85 111
8  189  90 120  93
9  172  95 110  86
10 170 125 130  78
```

  - 我们来作销售量 X_1 关于价格 X_2，广告费 X_3，以及有促销员 X_4 的一个线性回归拟合.

```
pullover.lm2 <- lm(X1 ~ ., data = pullover)
summary(pullover.lm2)
```

```
> summary(pullover.lm2)

Call:
lm(formula = X1 ~ ., data = pullover)

Residuals:
    Min      1Q  Median      3Q     Max
-13.369  -9.406   1.599   5.151  19.729

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 65.66956   57.12507   1.150  0.29407
X2          -0.21578    0.32194  -0.670  0.52764
X3           0.48519    0.08678   5.591  0.00139 **
X4           0.84373    0.37400   2.256  0.06491 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.7 on 6 degrees of freedom
Multiple R-squared:  0.9067     Adjusted R-squared:  0.8601
F-statistic: 19.44 on 3 and 6 DF,  p-value: 0.001713
```

$$X_1 = 65.66956 - 0.21578X_2 + 0.48519X_3 + 0.84373X_4$$

$$r^2 = 0.9067$$

X_1 的变化可由一个线性关系很好地近似.

多重线性模型 (Multiple Linear Model)

注 3.6 决定系数受解释变量个数的影响. 给定样本容量 n, r^2 会随着在线性模型中引入更多解释变量而增加. 因此, 即使引入可能不相关的解释变量, r^2 的值也可能会很高. 有 p 个解释变量以及含截距项 (共有 $p+1$ 个参数) 的调整后的决定系数为

$$r_{\text{adj}}^2 = r^2 - \frac{p\left(1-r^2\right)}{n-p-1}$$

多重线性模型 (Multiple Linear Model)

- 例：我们依然讨论套头衫数据集的例子.

```
pullover
```

```
> pullover
    X1  X2  X3  X4
1  230 125 200 109
2  181  99  55 107
3  165  97 105  98
4  150 115  85  71
5   97 120   0  82
6  192 100 150 103
7  181  80  85 111
8  189  90 120  93
9  172  95 110  86
10 170 125 130  78
```

  - 我们来作销售量 X_1 关于价格 X_2，广告费 X_3，以及有促销员 X_4 的一个线性回归拟合.

```
pullover.lm2 <- lm(X1 ~ ., data = pullover)
summary(pullover.lm2)
```

```
> summary(pullover.lm2)

Call:
lm(formula = X1 ~ ., data = pullover)

Residuals:
    Min      1Q  Median      3Q     Max
-13.369  -9.406   1.599   5.151  19.729

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 65.66956   57.12507   1.150  0.29407
X2          -0.21578    0.32194  -0.670  0.52764
X3           0.48519    0.08678   5.591  0.00139 **
X4           0.84373    0.37400   2.256  0.06491 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.7 on 6 degrees of freedom
Multiple R-squared:  0.9067     Adjusted R-squared:  0.8601
F-statistic: 19.44 on 3 and 6 DF,  p-value: 0.001713
```

$$X_1 = 65.66956 - 0.21578X_2 + 0.48519X_3 + 0.84373X_4$$

$r^2 = 0.9067$

X_1 的变化可由一个线性关系很好地近似.

$r^2_{\text{adj}} = 0.8601$

多重线性模型 (Multiple Linear Model)

- $\hat{\boldsymbol{\beta}}$ 的性质

$$\mathrm{E}\left(\hat{\boldsymbol{\beta}}\right)=\boldsymbol{\beta}$$

$$\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right)=\sigma^2\left(\mathcal{X}^{\mathrm{T}}\mathcal{X}\right)^{-1}$$

$$\widehat{\sigma^2}=\frac{1}{n-p-1}\left(\boldsymbol{y}-\hat{\boldsymbol{y}}\right)^{\mathrm{T}}\left(\boldsymbol{y}-\hat{\boldsymbol{y}}\right)$$

- 假设检验问题
  - 对假设 $\mathrm{H}_0:\ \beta_j=0$，检验统计量：$t=\dfrac{\widehat{\beta_j}}{\mathrm{SE}\left(\widehat{\beta_{\mathrm{j}}}\right)}$，当 $|t|>t_{n-p-1}\left(\dfrac{\alpha}{2}\right)$ 时拒绝 H_0

检验的显著水平

  - 对假设 $\mathrm{H}_0:\ \beta_1=\beta_2=\cdots=\beta_p=0$，我们用 F 统计量进行检验.

多重线性模型 (Multiple Linear Model)

- 例：我们依然讨论套头衫数据集的例子.

```
pullover
```

```
> pullover
    X1  X2  X3  X4
1  230 125 200 109
2  181  99  55 107
3  165  97 105  98
4  150 115  85  71
5   97 120   0  82
6  192 100 150 103
7  181  80  85 111
8  189  90 120  93
9  172  95 110  86
10 170 125 130  78
```

  - 我们来作销售量 X_1 关于价格 X_2，广告费 X_3，以及有促销员 X_4 的一个线性回归拟合.

```
pullover.lm2 <- lm(X1 ~ ., data = pullover)
summary(pullover.lm2)
```

```
> summary(pullover.lm2)

Call:
lm(formula = X1 ~ ., data = pullover)

Residuals:
    Min      1Q  Median      3Q     Max
-13.369  -9.406   1.599   5.151  19.729

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 65.66956   57.12507   1.150  0.29407
X2          -0.21578    0.32194  -0.670  0.52764
X3           0.48519    0.08678   5.591  0.00139 **
X4           0.84373    0.37400   2.256  0.06491 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.7 on 6 degrees of freedom
Multiple R-squared:  0.9067,    Adjusted R-squared:  0.8601
F-statistic: 19.44 on 3 and 6 DF,  p-value: 0.001713
```

$H_0: \beta_1 = 0 \leftrightarrow H_1: \beta_1 \neq 0 \Longrightarrow$ 接受 H_0

$H_0: \beta_2 = 0 \leftrightarrow H_1: \beta_2 \neq 0 \Longrightarrow$ 拒绝 H_0

$H_0: \beta_3 = 0 \leftrightarrow H_1: \beta_3 \neq 0 \Longrightarrow$ 拒绝 H_0

$$\begin{cases} H_0: \beta_1 = \beta_2 = \beta_3 = 0 \\ H_1: \text{at least one } \beta_j \neq 0 \end{cases} \Longrightarrow \text{拒绝 } H_0$$

多重线性模型 (Multiple Linear Model)

- 例：我们依然讨论套头衫数据集的例子.

```
pullover
```

```
> pullover
    X1  X2  X3  X4
1  230 125 200 109
2  181  99  55 107
3  165  97 105  98
4  150 115  85  71
5   97 120   0  82
6  192 100 150 103
7  181  80  85 111
8  189  90 120  93
9  172  95 110  86
10 170 125 130  78
```

  - 我们来作销售量 X_1 关于价格 X_2，广告费 X_3，以及有促销员 X_4 的一个线性回归拟合.

```
pullover.lm2 <- lm(X1 ~ ., data = pullover)
summary(pullover.lm2)
```

```
> summary(pullover.lm2)

Call:
lm(formula = X1 ~ ., data = pullover)

Residuals:
    Min      1Q  Median      3Q     Max
-13.369  -9.406   1.599   5.151  19.729

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 65.66956   57.12507   1.150  0.29407
X2          -0.21578    0.32194  -0.670  0.52764
X3           0.48519    0.08678   5.591  0.00139 **
X4           0.84373    0.37400   2.256  0.06491 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.7 on 6 degrees of freedom
Multiple R-squared:  0.9067,    Adjusted R-squared:  0.8601
F-statistic: 19.44 on 3 and 6 DF,  p-value: 0.001713
```

$H_0: \beta_1 = 0 \leftrightarrow H_1: \beta_1 \neq 0 \Longrightarrow$ 接受 H_0

$H_0: \beta_2 = 0 \leftrightarrow H_1: \beta_2 \neq 0 \Longrightarrow$ 拒绝 H_0

$H_0: \beta_3 = 0 \leftrightarrow H_1: \beta_3 \neq 0 \Longrightarrow$ 拒绝 H_0

$$\begin{cases} H_0: \beta_1 = \beta_2 = \beta_3 = 0 \\ H_1: \text{at least one } \beta_j \neq 0 \end{cases} \Longrightarrow \text{拒绝 } H_0$$

Boston 住房数据集

- Boston 住房数据集由 Boston 大区的每个人口普查地区的共 506 个观测值构成.

```
library(MASS)
str(Boston)
```

```
> str(Boston)    506 个观测值      14 个变量
'data.frame':   506 obs. of  14 variables:
 $ crim    : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
 $ zn      : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
 $ indus   : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
 $ chas    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ nox     : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
 $ rm      : num  6.58 6.42 7.18 7 7.15 ...
 $ age     : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
 $ dis     : num  4.09 4.97 4.97 6.06 6.06 ...
 $ rad     : int  1 2 2 3 3 3 5 5 5 5 ...
 $ tax     : num  296 242 242 222 222 222 311 311 311 311 ...
 $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
 $ black   : num  397 397 393 395 397 ...
 $ lstat   : num  4.98 9.14 4.03 2.94 5.33 ...
 $ medv    : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

Boston 住房数据集

- Boston 住房数据集由 Boston 大区的每个人口普查地区的共 506 个观测值构成.

```
head(Boston)
```

X_{14}：自住房屋价格的中位数(单位: 1000美元)

X_6：每个住宅的平均房间数

X_4：Charles 河 (1 指河边，0 为其它)

X_8：到波士顿五个就业中心的加权距离

X_{10}：每一万美元的全额财产税税率

```
> head(Boston)
     crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat medv
1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98 24.0
2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14 21.6
3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03 34.7
4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94 33.4
5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33 36.2
6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21 28.7
```

X_1：人均犯罪率

X_2：划作大型住宅用地的比例

X_3：非零售业务占地的比例

X_5：一氧化氮浓度

X_7：1940年之前建造的自住房的比例

X_9：辐射状高速公路的可达性指数

X_{11}：生师比

X_{12}：$1000\left(B-0.63\right)^2 \mathrm{I}(B<0.63)$

其中 B是非裔美国人的比例

X_{13}：社会低层人口的百分比

Boston 住房数据集

- 描述性统计

```
Boston_mean <- sapply(Boston, mean)
Boston_median <- sapply(Boston, median)
Boston_var <- sapply(Boston, var)
Boston_sd <- sapply(Boston, sd)
Boston_descriptive <- tibble(Mean = Boston_mean, Median = Boston_median, Variance = Boston_var,
                             Std = Boston_sd)
Boston_descriptive
```

```
> Boston_descriptive
# A tibble: 14 × 4
      Mean  Median   Variance     Std
     <dbl>   <dbl>      <dbl>   <dbl>
 1   3.61    0.257    74.0      8.60
 2  11.4     0       544.      23.3
 3  11.1     9.69     47.1      6.86
 4   0.0692  0         0.0645   0.254
 5   0.555   0.538     0.0134   0.116
 6   6.28    6.21      0.494    0.703
 7  68.6    77.5     792.      28.1
 8   3.80    3.21      4.43     2.11
 9   9.55    5        75.8      8.71
10 408.    330     28405.    169.
11  18.5    19.0      4.69     2.16
12 357.    391.     8335.     91.3
13  12.7    11.4     51.0      7.14
14  22.5    21.2     84.6      9.20
```

Boston 住房数据集

- 样本协方差矩阵的无偏估计

```
S <- cov(Boston)
round(S, digits = 2)
```

```
> round(S, digits = 2)
           crim       zn   indus  chas   nox     rm      age      dis      rad      tax ptratio    black   lstat    medv
crim      73.99   -40.22   23.99 -0.12  0.42  -1.33    85.41    -6.88    46.85   844.82    5.40  -302.38   27.99  -30.72
zn       -40.22   543.94  -85.41 -0.25 -1.40   5.11  -373.90    32.63   -63.35 -1236.45  -19.78   373.72  -68.78   77.32
indus     23.99   -85.41   47.06  0.11  0.61  -1.89   124.51   -10.23    35.55   833.36    5.69  -223.58   29.58  -30.52
chas      -0.12    -0.25    0.11  0.06  0.00   0.02     0.62    -0.05    -0.02    -1.52   -0.07     1.13   -0.10    0.41
nox        0.42    -1.40    0.61  0.00  0.01  -0.02     2.39    -0.19     0.62    13.05    0.05    -4.02    0.49   -0.46
rm        -1.33     5.11   -1.89  0.02 -0.02   0.49    -4.75     0.30    -1.28   -34.58   -0.54     8.22   -3.08    4.49
age       85.41  -373.90  124.51  0.62  2.39  -4.75   792.36   -44.33   111.77  2402.69   15.94  -702.94  121.08  -97.59
dis       -6.88    32.63  -10.23 -0.05 -0.19   0.30   -44.33     4.43    -9.07  -189.66   -1.06    56.04   -7.47    4.84
rad       46.85   -63.35   35.55 -0.02  0.62  -1.28   111.77    -9.07    75.82  1335.76    8.76  -353.28   30.39  -30.56
tax      844.82 -1236.45  833.36 -1.52 13.05 -34.58  2402.69  -189.66  1335.76 28404.76  168.15 -6797.91  654.71 -726.26
ptratio    5.40   -19.78    5.69 -0.07  0.05  -0.54    15.94    -1.06     8.76   168.15    4.69   -35.06    5.78  -10.11
black   -302.38   373.72 -223.58  1.13 -4.02   8.22  -702.94    56.04  -353.28 -6797.91  -35.06  8334.75 -238.67  279.99
lstat     27.99   -68.78   29.58 -0.10  0.49  -3.08   121.08    -7.47    30.39   654.71    5.78  -238.67   50.99  -48.45
medv     -30.72    77.32  -30.52  0.41 -0.46   4.49   -97.59     4.84   -30.56  -726.26  -10.11   279.99  -48.45   84.59
```

Boston 住房数据集

- 样本相关矩阵

```
R <- cor(Boston)
round(R, digits = 2)
```

```
> R <- cor(Boston)
> round(R, digits = 2)
        crim    zn indus  chas   nox    rm   age   dis   rad   tax ptratio black lstat  medv
crim    1.00 -0.20  0.41 -0.06  0.42 -0.22  0.35 -0.38  0.63  0.58    0.29 -0.39  0.46 -0.39
zn     -0.20  1.00 -0.53 -0.04 -0.52  0.31 -0.57  0.66 -0.31 -0.31   -0.39  0.18 -0.41  0.36
indus   0.41 -0.53  1.00  0.06  0.76 -0.39  0.64 -0.71  0.60  0.72    0.38 -0.36  0.60 -0.48
chas   -0.06 -0.04  0.06  1.00  0.09  0.09  0.09 -0.10 -0.01 -0.04   -0.12  0.05 -0.05  0.18
nox     0.42 -0.52  0.76  0.09  1.00 -0.30  0.73 -0.77  0.61  0.67    0.19 -0.38  0.59 -0.43
rm     -0.22  0.31 -0.39  0.09 -0.30  1.00 -0.24  0.21 -0.21 -0.29   -0.36  0.13 -0.61  0.70
age     0.35 -0.57  0.64  0.09  0.73 -0.24  1.00 -0.75  0.46  0.51    0.26 -0.27  0.60 -0.38
dis    -0.38  0.66 -0.71 -0.10 -0.77  0.21 -0.75  1.00 -0.49 -0.53   -0.23  0.29 -0.50  0.25
rad     0.63 -0.31  0.60 -0.01  0.61 -0.21  0.46 -0.49  1.00  0.91    0.46 -0.44  0.49 -0.38
tax     0.58 -0.31  0.72 -0.04  0.67 -0.29  0.51 -0.53  0.91  1.00    0.46 -0.44  0.54 -0.47
ptratio 0.29 -0.39  0.38 -0.12  0.19 -0.36  0.26 -0.23  0.46  0.46    1.00 -0.18  0.37 -0.51
black  -0.39  0.18 -0.36  0.05 -0.38  0.13 -0.27  0.29 -0.44 -0.44   -0.18  1.00 -0.37  0.33
lstat   0.46 -0.41  0.60 -0.05  0.59 -0.61  0.60 -0.50  0.49  0.54    0.37 -0.37  1.00 -0.74
medv   -0.39  0.36 -0.48  0.18 -0.43  0.70 -0.38  0.25 -0.38 -0.47   -0.51  0.33 -0.74  1.00
```

$X_1 \rightarrow$ crim
$X_2 \rightarrow$ zn
$X_3 \rightarrow$ indus
$X_4 \rightarrow$ chas
$X_5 \rightarrow$ nox
$X_6 \rightarrow$ rm
$X_7 \rightarrow$ age
$X_8 \rightarrow$ dis
$X_9 \rightarrow$ rad
$X_{10} \rightarrow$ tax
$X_{11} \rightarrow$ ptratio
$X_{12} \rightarrow$ black
$X_{13} \rightarrow$ lstat
$X_{14} \rightarrow$ medv

X_6 X_{10} X_{11} X_{13}

变量 X_{14} 与其它变量的相关系数

Boston 住房数据集

- 对假设 H_0： $\rho_{X_{14}X_i} = 0$， $i = 1, 2, \ldots, 13$，应用 Fisher 的 Z 变换

$$W = \frac{1}{2}\log\left(\frac{1+r_{XY}}{1-r_{XY}}\right)$$

$$\mathrm{E}(W) \approx \frac{1}{2}\log\left(\frac{1+\rho_{XY}}{1-\rho_{XY}}\right), \qquad \mathrm{Var}(W) \approx \frac{1}{n-3}$$

```
R_14 <- R[14, 1:13]
W_14 <- 0.5 * log((1 + R_14) / (1 - R_14))
Z_14 <- W_14 / sqrt(1/(dim(Boston)[1]-3))
p_value <- 2 * pnorm(abs(Z_14), lower.tail = FALSE)
p_value
```

拒绝 H_0： $\rho_{X_{14}X_i} = 0$

```
> p_value
          crim             zn          indus           chas            nox             rm            age
 3.895604e-20  2.580349e-17  2.479294e-32  7.136774e-05  1.301673e-24  1.445868e-82  5.965212e-19
           dis            rad            tax        ptratio          black          lstat
 1.024590e-08  1.967604e-19  4.299209e-30  3.792391e-36  7.482300e-15 9.230431e-100
```

- 相关系数与 Fisher 的 Z 变换不适用于二值变量，所以，上述检验结果不适用于变量 X_{14} 与 X_4.

Boston 住房数据集

- 如果我们希望用变量 $X_1, X_2, \ldots, X_{13}$ 来解释价格 X_{14} 的变化，我们可以拟合下述线性模型

$$X_{14} = \beta_0 + \sum_{j=1}^{13} \beta_j X_j + \varepsilon$$

```
Boston.lm <- lm(medv ~ ., data = Boston)
summary(Boston.lm)
```

```
> summary(Boston.lm)

Call:
lm(formula = medv ~ ., data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-15.595  -2.730  -0.518   1.777  26.199

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
zn           4.642e-02  1.373e-02   3.382 0.000778 ***
indus        2.056e-02  6.150e-02   0.334 0.738288
chas         2.687e+00  8.616e-01   3.118 0.001925 **
nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
age          6.922e-04  1.321e-02   0.052 0.958229
dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
black        9.312e-03  2.686e-03   3.467 0.000573 ***
lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

对 X_{14} 的影响不显著

$r^2 = 0.7406$

$r^2_{\text{adj}} = 0.7338$