第1章 基础知识

- **1.1** 探索性数据分析
- **1.2** 概率统计
- **1.3** 矩阵代数
- **1.4** 统计图形

1.1 探索性数据分析

- 探索性数据分析
- Exploratory Data Analysis（EDA）
- 是对调查、观察所得到的一些初步的杂乱无章的数据，在尽量少的先验假定下进行处理，通过作图、制表等形式和方程拟合、计算某些特征量等手段，探索数据的结构和规律的一种数据分析方法。

David C. Hoaglin等著，陈忠琏等译.探索性数据分析.北京：中国统计出版社，1998

1.1 探索性数据分析

- 探索性数据分析强调灵活探求线索和证据；而证实性数据分析则着重评估现有证据。
- 无论是对一大组数据，还是对相继的几小组数据作分析，一般都要经过探索和证实这两个阶段。
- 通常还要交替的使用探索性技术和证实性技术，循环反复多次，才能得到满意的结果。

1.1 探索性数据分析

EDA的主要工作：

- 对数据进行清洗，

- 对数据进行描述（描述统计量，图表），查看数据的分布，

- 比较数据之间的关系，培养对数据的直觉，

- 对数据进行总结等。

探索性数据分析的四大主题

1、耐抗性(Resistnace)

即对数据的不良表现(如极端值或称奇异点)不敏感，也就是说对于数据的任意一个小部分的很大的改变，或者对于数据的大部分的很小改变，(统计)分析或概括仅产生很小的变化。

1.1 探索性数据分析

2、残差(Residuals)

残差是从原始数据中减去概括性统计量或所配合模型的趋势值后所剩余的部分。其公式为:

残差=原数据-拟合值

1.1 探索性数据分析

3、重新表述(Re-expression)

重新表达(Re-expression)，涉及到运用何种尺度会简化分析。

1.1 探索性数据分析

4、图形启示(Revelation)

探索性数据分析强调数据图形的启示作用，它能使分析者看出数据、拟合以及残差的行为，从而抓住数据中意想不到的特点。

【例】Bendixen（1977）给出了需要24小时以上呼吸支持（一种强化治疗）的11类病人的生存百分率。分析什么百分率是典型的。

次序统计量为

i：	1	2	3	4	5	6	7	8	9	10	11
$x_{(i)}$：	36	37	45	52	56	58	66	68	75	90	100

由于$n=11$，中位数深度$d(M)=(11+1)/2=6$，中位数$M=x_{(6)}=58$；四分数深度$d(F)=(6+1)/2=3.5$，因而下四分数$F_l=(x_{(3)}+x_{(4)})/2=48.5$，上四分数$F_u=(x_{(9)}+x_{(8)})/2=71.5$

将中位数、极端数、四分数放在一起的五数总括可知：这11类病人生存百分率的典型值是58%，尽管生存率可以高达100%，低到36%，但其中一半的生存率是48.5%~71.5%

1.2 概率统计

- 统计必须为各个领域服务
- 统计必须和数据打交道
- 统计必须和计算机结合

1.2 概率统计：统计量

1，位置特征量

- 均值
- 中位数
- 众数

2，分散程度（变异性）

- 方差和标准差
- 变异系数
- 极差
- 四分位间距

3，与分布形状有关

- 偏度
- 峰度

4，相关系数

1.2 概率统计

• 随机过程

设 T 是一无限实数集。我们把依赖于参数 $t \in T$ 的一族（无限多个）随机变量称为**随机过程**，记为 $\{X(t), t \in T\}$ 。

- $X(t)$ 是随机变量， T 叫做参数集。
- $X(t_1) = x$ 称为 t_1 时刻的**状态**，
- 全体 X(t) 的可能取值称为随机过程的**状态空间**。

1.2 概率统计

- **两类随机过程：**

1，到达过程

关注的是某种“到达”的事件是否发生

2，马尔科夫过程

未来的数据只依赖于当前的数据，而与过去的历史数据无关

1.3 矩阵代数

1 三角分解，正交-三角分解

2 奇异值分解

3 矩阵的广义逆

1 正交-三角分解

- **QR分解**或**正交-三角分解**

对矩阵A，

若有矩阵Q满足$Q^TQ = I$，

以及上三角矩阵R使得

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

则称为矩阵A的QR分解

正交-三角分解----Schimidt正交化方法

为了叙述方便，假定矩阵A非奇异。下面以n=3为例进行说明。设$\alpha_1, \alpha_2, \alpha_3$, 是A的各列构成的列向量，它们线性无关。

令$\beta_1' = \alpha_1$, $\beta_1 = \beta_1' / \|\beta_1'\|_2$, 则β_1是一个单位长度的向量。再令

$$\beta_2' = \alpha_2 - (\alpha_2, \beta_1)\,\beta_1, \qquad \beta_2 = \beta_2' / \|\beta_2'\|_2 ,$$

可以验证$(\beta_1, \beta_2) = 0$,即β_1和β_2正交。

正交-三角分解----Schimidt正交化方法

进一步，若令

$$\beta_3' = \alpha_3 - (\alpha_3, \beta_1)\,\beta_1 - (\alpha_3, \beta_2)\,\beta_2,$$

则

$$(\beta_3', \beta_1) = (\beta_3', \beta_2) = 0,$$

即β_3'和β_1、β_2正交。将其单位化

$$\beta_3 = \beta_3' / \|\beta_3'\|_2 \,,$$

正交-三角分解----Schimidt正交化方法

于是

$$A=[\alpha_1,\alpha_2,\alpha_3]=[\beta_1,\beta_2,\beta_3]\begin{bmatrix} \|\beta_1'\|_2 & (\alpha_2,\beta_1) & (\alpha_3,\beta_1) \\ & \|\beta_2'\|_2 & (\alpha_3,\beta_2) \\ & & \|\beta_3'\|_2 \end{bmatrix}=QR$$

其中，Q=[β_1, β_2 , β_3]是正交矩阵，R是上三角矩阵。

正交-三角分解----Schimidt正交化方法

设A为n阶非奇异矩阵，$\alpha_1, \alpha_2, \ldots, \alpha_n$, 是A的各列构成的线性无关的向量。类似地有：

$\beta_1' = \alpha_1$,　　　　$\beta_1 = \beta_1' / \|\beta_1'\|_2$,

$\beta_2' = \alpha_2 - (\alpha_2, \beta_1)\, \beta_1$,　　　　$\beta_2 = \beta_2' / \|\beta_2'\|_2$

$\beta_3' = \alpha_3 - (\alpha_3, \beta_1)\, \beta_1 - (\alpha_3, \beta_2)\, \beta_2$,　　　　$\beta_3 = \beta_3' / \|\beta_3'\|_2$

……

$\beta_n = \beta_n' / \|\beta_n'\|_2$,

可以验证$(\beta_i, \beta_j) = 0$ $(i \neq j)$,即β_i和β_j正交。

$$\beta_n' = \alpha_n - \sum_{j=1}^{n-1} (\alpha_n, \beta_j)\beta_j$$

正交-三角分解----Schimidt正交化方法

于是 $A=[\alpha_1,\alpha_2,\cdots,\alpha_n]$

$$
=[\beta_1,\beta_2,\cdots,\beta_n]\begin{bmatrix} \|\beta_1'\|_2 & (\alpha_2,\beta_1) & (\alpha_3,\beta_1) & \cdots & (\alpha_n,\beta_1) \\ & \|\beta_2'\|_2 & (\alpha_3,\beta_2) & \cdots & (\alpha_n,\beta_2) \\ & & \|\beta_3'\|_2 & \cdots & (\alpha_n,\beta_3) \\ & & & \ddots & \vdots \\ & & & & \|\beta_n'\|_2 \end{bmatrix}
$$

$$
=QR
$$

其中，Q是正交矩阵，R是上三角矩阵。此方法称为Schimidt正交化方法。

正交-三角分解----Schimidt正交化方法

例 设矩阵

$$A=\begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 4 \end{bmatrix}$$

试用Schimidt正交化方法分解

正交-三角分解----Schimidt正交化方法

解：

$\alpha_1 = [2,1,0]^T$

$\alpha_2 = [1,2,1]^T$

$\alpha_3 = [0,1,4]^T$

$\beta_1' = \alpha_1$

$\|\beta_1'\|_2 = \sqrt{2^2 + 1^2} = \sqrt{5} = 2.236$

$\beta_1 = \beta_1'/\|\beta_1'\|_2 = [0.89443,0.44721,0]^T$

$\beta_2' = \alpha_2 - (\alpha_2, \beta_1)\beta_1 = [-0.6,1.2,1]^T$

$\beta_2 = \beta_2'/\|\beta_2'\|_2 = [-0.3585,0.7171,0.5976]^T$

$\beta_3' = \alpha_3 - (\alpha_3, \beta_1)\beta_1 - (\alpha_3, \beta_2)\beta_2 = [0.71429, -1.42857,2.14286]^T$

$\beta_3 = \beta_3'/\|\beta_3'\|_2 = [-0.22450,0.44900,0.37417]^T$

$Q = (\beta_1, \beta_2, \beta_3)$

$$R = \begin{pmatrix} 2.23607 & 1.7885 & 0.44721 \\ 0 & 1.67332 & 3.10759 \\ 0 & 0 & 2.6726 \end{pmatrix}$$

2 奇异值分解（SVD）

1 三角分解，正交-三角分解

2 奇异值分解

3 矩阵的广义逆

方阵对角化变换（特征值分解）

- 如果线性变换的特征向量可以构成线性空间的一组基，那么有 $A = Q\Sigma Q^{-1}$

其中Q为特征向量组成的矩阵，Σ为对角阵，对角线上的元素为特征向量对应的特征值。

表示线性变换（矩阵）可以由其特征根和特征向量还原

简证：

$$Q = \begin{bmatrix} V_1 & V_2 & V_3 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}$$

$$Q\Sigma = \begin{bmatrix} \lambda_1 V_1 & \lambda_2 V_2 & \lambda_3 V_3 \end{bmatrix} = \begin{bmatrix} AV_1 & AV_2 & AV_3 \end{bmatrix} = AQ$$

$$A = AQQ^{-1} = Q\Sigma Q^{-1}$$

方阵还原

- 矩阵维度相当大时，可以用对角化变化的方法来逼近矩阵：

特征向量：矩阵的特征

特征值：各特征所占的权重

$$A = Q\Sigma Q^{-1}$$

$$A = \left(v_1, v_2, \cdots, v_n\right)\begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix}\begin{pmatrix} v_1' \\ \vdots \\ v_n' \end{pmatrix}$$

$$= \lambda_1 v_1 v_1' + \lambda_2 v_2 v_2' + \cdots + \lambda_n v_n v_n'$$

$$\approx \lambda_1 v_1 v_1' + \lambda_2 v_2 v_2' + \cdots + \lambda_r v_r v_r' \quad , \qquad r \le n$$

思考?

- 特征值分解使用条件有限，仅适用于**方阵**，且要求是**实对称**矩阵。

- 试想：如果需要分析的矩阵不是方阵（事实经常如此），该怎样处理呢？

SVD简介

- SVD：Singular Value Decomposition，奇异值分解
- SVD可以看做是特征值分解的一种推广
- 或者说特征值分解可以看作是SVD的一种特例
- 当矩阵不是方阵时同样适用，应用很广

SVD分解(1)

对任意矩阵$A \in R^{m\times n}$，$rank(A)=r$,总可以取A 的如下分解

左奇异向量

右奇异向量

$$A = U\Sigma_A V^T$$

非零奇异值，**从大到小**依次排序

其中U、V为正交矩阵

$$\Sigma_A = \begin{pmatrix} \sigma_1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \sigma_r & 0 \\ 0 & \cdots & 0 & 0 \end{pmatrix}, \quad \left(A^T A\right) v_i = \lambda_i v_i, \quad \sigma_i = \sqrt{\lambda_i}$$

SVD分解(2)

奇异值从大到小
依次排序

SVD分解(2)

$A = U\Sigma_A V^T$

$\begin{cases} AV = U\Sigma_A V^T V = U\Sigma_A, \\ U^T A = U^T U \Sigma_A V^T = \Sigma_A V^T \end{cases}, \quad \begin{cases} U = (u_1, u_2, \cdots, u_m), \quad u_i: m \times 1\text{向量} \\ V = (v_1, v_2, \cdots, v_n), \quad v_i: n \times 1\text{向量} \end{cases}$

$$AV = U\Sigma_A \Rightarrow (Av_1, \cdots, Av_n) = (u_1, \cdots, u_m)\Sigma_A$$

$$\Rightarrow (Av_1, \cdots, Av_n) = (u_1, \cdots, u_m)\begin{pmatrix} \sigma_1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \sigma_r & 0 \\ 0 & \cdots & 0 & 0 \end{pmatrix}$$

$$= (u_1\sigma_1, \cdots, u_\mathrm{r}\sigma_r, 0)$$

$$\begin{cases} Av_1 = u_1\sigma_1 \\ \cdots \\ Av_r = u_r\sigma_r \end{cases}$$

奇异值从大到小依次排序

SVD分解(3)

$$U^T A = \Sigma_A V^T$$

$$\begin{cases} U^T A = \begin{pmatrix} u_1^T \\ \vdots \\ u_m^T \end{pmatrix} A = \begin{pmatrix} u_1^T A \\ \vdots \\ u_m^T A \end{pmatrix} \\ \Sigma_A V^T = \begin{pmatrix} \sigma_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \sigma_r & 0 \\ 0 & \cdots & 0 & 0 \end{pmatrix} \begin{pmatrix} v_1^T \\ \vdots \\ v_n^T \end{pmatrix} = \begin{pmatrix} \sigma_1 v_1^T \\ \vdots \\ \sigma_r v_r^T \\ 0 \end{pmatrix} \end{cases}$$

奇异值**从大到小**依次排序

$$\Rightarrow \begin{pmatrix} u_1^T A \\ \vdots \\ u_m^T A \end{pmatrix} = \begin{pmatrix} \sigma_1 v_1^T \\ \vdots \\ \sigma_r v_r^T \\ 0 \end{pmatrix} \Rightarrow \begin{cases} u_1^T A = \sigma_1 v_1^T \\ \vdots \\ u_r^T A = \sigma_r v_r^T \end{cases}$$

SVD算法解析

$$\Sigma_A = \begin{pmatrix} \sigma_1 & \square & 0 & 0 \\ \square & \square & \square & \square \\ 0 & \square & \sigma_r & 0 \\ 0 & \square & 0 & 0 \end{pmatrix},\qquad \left(A^T A\right)v_i = \lambda_i v_i,\quad \sigma_i = \sqrt{\lambda_i}$$

奇异值从大到小
依次排序

可以看作是矩阵A的“平方”，而奇异值又是A“平方”阵的特征根的开方，因此**奇异值**可以看作是矩阵A的**“伪特征值”**，**左奇异向量**可以看作矩阵A的**“行特征向量”**，**右奇异向量**可以看作是矩阵A的**“列特征向量”**。

SVD矩阵近似（1）

$$A = U\Sigma_A V^T,$$

$U = (u_1, u_2, \cdots, u_m), u_i : m \times 1$向量（代表矩阵A的行特征）

$V = (v_1, v_2, \cdots, v_n), v_i : n \times 1$向量（代表矩阵A的列特征）

$$\Sigma_A = \begin{pmatrix} \sigma_1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \sigma_r & 0 \\ 0 & \cdots & 0 & 0 \end{pmatrix},$$

奇异值从大到小依次排序

$$A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_r u_r v_r^T$$

SVD矩阵近似（2）

$$A=\sigma_1 u_1 v_1^T+\sigma_2 u_2 v_2^T+\cdots+\sigma_r u_r v_r^T$$

σ_1，$\sigma_2,\cdots$，σ_r为矩阵A的非零奇异值，从大到小依次排列，且下降的非常快。奇异值代表的对应的“行列特征”的重要程度。

因此可以提取矩阵的前t 个特征来近似矩阵：

$$A\approx\sigma_1 u_1 v_1^T+\sigma_2 u_2 v_2^T+\cdots+\sigma_t u_t v_t^T$$

• SVD应用—推荐算法

会员		电影							
		喜剧				恐怖			
偏好	ID	宿醉	东成西就	大话西游	八星报喜	午夜凶铃	咒怨	林中小屋	寂静岭
喜剧	至尊宝	4	4	5	5	2	3	2	3.75
	小小宝	5	5	5	4	2	2	3	1
	流氓兔	5	4	4	5	2	3	1	2
	霹*雳	5	4	5	5	3	2	1	2
	中原不败	4	5	5	4	2	1	3	2
恐怖	魂飞魄散	1	2	3	2	5	3.875	5	5
	荒村少年	3	1	2	2	4	5	4	4
	憨豆豆	2	1	3	2	4	5	4	5
	怪大叔	2	2	3	1	5	5	5	4
	美味僵尸	1	3	2	1	4	5	4	5

SVD——矩阵变换

矩阵$A(10\times 8)$

$$\begin{bmatrix} 4 & 4 & 5 & 5 & 2 & 3 & 2 & 3.75 \\ 5 & 5 & 5 & 4 & 2 & 2 & 3 & 1 \\ 5 & 4 & 4 & 5 & 2 & 3 & 1 & 2 \\ 5 & 4 & 5 & 5 & 3 & 2 & 1 & 2 \\ 4 & 5 & 5 & 4 & 2 & 1 & 3 & 2 \\ 1 & 2 & 3 & 2 & 5 & 3.875 & 5 & 5 \\ 3 & 1 & 2 & 2 & 4 & 5 & 4 & 4 \\ 2 & 1 & 3 & 2 & 4 & 5 & 4 & 5 \\ 2 & 2 & 3 & 1 & 5 & 5 & 5 & 4 \\ 1 & 3 & 2 & 1 & 4 & 5 & 4 & 5 \end{bmatrix}$$

$$\begin{bmatrix} 4 & 5 & 5 & 5 & 4 & 1 & 3 & 2 & 2 & 1 \\ 4 & 5 & 4 & 4 & 5 & 2 & 1 & 1 & 2 & 3 \\ 5 & 5 & 4 & 5 & 5 & 3 & 2 & 3 & 3 & 2 \\ 5 & 4 & 5 & 5 & 4 & 2 & 2 & 2 & 1 & 1 \\ 2 & 2 & 2 & 3 & 2 & 5 & 4 & 4 & 5 & 4 \\ 3 & 2 & 3 & 2 & 1 & 3.875 & 5 & 5 & 5 & 5 \\ 2 & 3 & 1 & 1 & 3 & 5 & 4 & 4 & 5 & 4 \\ 3.75 & 1 & 2 & 2 & 2 & 5 & 4 & 5 & 4 & 5 \end{bmatrix}$$

矩阵$A^T\ (8\times 10)$

$$A^T A = \left(A^T A\right) = \begin{bmatrix} 126 & 115 & 133 & 121 & 90 & 95 & 84 & 88 \\ 115 & 117 & 129 & 113 & 88 & 90 & 86 & 88 \\ 133 & 129 & 151 & 131 & 111 & 114 & 107 & 112 \\ 121 & 113 & 131 & 121 & 86 & 90 & 79 & 88 \\ 90 & 88 & 111 & 86 & 123 & 128 & 119 & 125 \\ 95 & 90 & 114 & 90 & 128 & 142 & 124 & 135 \\ 84 & 86 & 107 & 79 & 119 & 124 & 122 & 122 \\ 88 & 88 & 112 & 88 & 125 & 135 & 122 & 134 \end{bmatrix}$$

8×8矩阵

SVD——求奇异值

矩阵$\left(A^T A\right)$的非零特征值为{879.7,130.79,12.17,6.4,3.9},对应矩阵A 的非零奇异值为{$\sigma_1=29.7$, $\sigma_2=11.4$, $\sigma_3=3.5$, $\sigma_4=2.5$},对应右奇异向量为:

$$v_1=\begin{pmatrix}0.34\\0.33\\0.40\\0.33\\0.35\\0.37\\0.34\\0.36\end{pmatrix},\quad v_2=\begin{pmatrix}0.39\\0.34\\0.28\\0.40\\-0.31\\-0.36\\-0.34\\-0.36\end{pmatrix}$$

由于奇异值（特征的权重）下降的速度非常快，表明矩阵的信息量集中分布在前几个较大的特征值中，本例中提取前2个特征。

SVD——右奇异向量解析

影片类型	片名	特征1（29.7）	特征2（11.4）	得分均值
喜剧	宿醉	0.34	0.39	3.20
	东成西就	0.33	0.34	3.10
	大话西游	0.40	0.29	3.70
	八星报喜	0.33	0.40	3.10
恐怖	午夜凶铃	0.35	-0.31	3.30
	咒怨	0.37	-0.37	3.49
	林中小屋	0.34	-0.34	3.20
	寂静岭	0.36	-0.37	3.38

可以看作电影的本身的**精彩程度**的特征

可以看做有关电影**影片类型**的特征

SVD——左奇异向量解析

偏好	ID	特征1（29.7）	特征2（11.4）	打分平均值
喜剧	至尊宝	0.34	0.23	3.59
	小小宝	0.32	0.34	3.38
	流氓兔	0.31	0.32	3.25
	霹*雳	0.32	0.35	3.38
	中原不败	0.31	0.31	3.25
恐怖	魂飞魄散	0.32	-0.33	3.36
	荒村少年	0.30	-0.27	3.13
	憨豆豆	0.31	-0.31	3.25
	怪大叔	0.32	-0.34	3.38
	美味僵尸	0.30	-0.34	3.13

可以看做是会员的**打分习惯**特征

可看做是会员对影片**类型偏好**的特征

SVD——模型打分（1）

至尊宝	0.34	0.23
小小宝	0.32	0.34
流氓兔	0.31	0.32
霹*雳	0.32	0.35
中原不败	0.31	0.31
魂飞魄散	0.32	−0.33
荒村少年	0.30	−0.27
憨豆豆	0.31	−0.31
怪大叔	0.32	−0.34
美味僵尸	0.30	−0.34

$\cdot \begin{pmatrix} 29.7 & \\ & 11.4 \end{pmatrix} \cdot$

宿醉	东成西就	大话西游	八星报喜	午夜凶铃	咒怨	林中小屋	寂静岭
0.34	0.33	0.40	0.33	0.35	0.37	0.34	0.36
0.39	0.34	0.29	0.40	−0.31	−0.37	−0.34	−0.37

$\approx A$

各部分得分权重

$$29.7\times(0.34\times0.36)+11.4\times(0.23\times-0.37)=2.6$$

计算至尊宝对《寂静岭》的评分

影片相对客观分数

影片类型适应度得分得分

SVD——模型打分（2）

$$
\begin{array}{l}
\text{至尊宝}\\ \text{小小宝}\\ \text{流氓兔}\\ \text{霹*雳}\\ \text{中原不败}\\ \text{魂飞魄散}\\ \text{荒村少年}\\ \text{憨豆豆}\\ \text{怪大叔}\\ \text{美味僵尸}
\end{array}
\begin{pmatrix}
0.34 & 0.23\\
0.32 & 0.34\\
0.31 & 0.32\\
0.32 & 0.35\\
0.31 & 0.31\\
0.32 & -0.33\\
0.30 & -0.27\\
0.31 & -0.31\\
0.32 & -0.34\\
0.30 & -0.34
\end{pmatrix}
\cdot
\begin{pmatrix}
29.7 & \\
 & 11.4
\end{pmatrix}
\cdot
\begin{array}{cccccccc}
\text{宿醉} & \text{东成西就} & \text{大话西游} & \text{八星报喜} & \text{午夜凶铃} & \text{咒怨} & \text{林中小屋} & \text{寂静岭}\\
\end{array}
\begin{pmatrix}
0.34 & 0.33 & 0.40 & 0.33 & 0.35 & 0.37 & 0.34 & 0.36\\
0.39 & 0.34 & 0.29 & 0.40 & -0.31 & -0.37 & -0.34 & -0.37
\end{pmatrix}
\approx A
$$

各部分得分权重

$$29.7\times(0.32\times 0.37)+11.4\times(-0.33\times -0.37)=4.9$$

计算魂飞魄散对《咒怨》的评分

影片相对客观分数

影片类型适应度得分得分

SVD结果简要测评

- 至尊宝的观影历史显示其对喜剧类的电影评分较高，对恐怖电影普遍评分较低，因此可以推测他应该是不喜欢看《寂静岭》的，模型给出的打分为**2.6**，与实际情况是相符的。
- 魂飞魄散的观影历史显示其对恐怖类的电影评分较高，对喜剧电影普遍评分较低，因此可以推测他应该是很喜欢看《咒怨》的，模型给出的打分为**4.9**，与实际情况是相符的。

偏好	ID	宿醉	东成西就	大话西游	八星报喜	午夜凶铃	咒怨	林中小屋	寂静岭
喜剧	至尊宝	4	4	5	5	2	3	2	2.6
恐怖	魂飞魄散	1	2	3	2	5	4.9	5	5

3 矩阵的广义逆

3.1 广义逆的定义和性质

3.2 叉积阵

3.3 正交投影

3.1 广义逆的定义和性质

定义（加号逆和减号逆）设A为(n, m)矩阵， 若(m, n)矩阵G满足

$$
\begin{aligned}
&\text{i)} && AGA = A; \\
&\text{ii)} && GAG = G; \\
&\text{iii)} && (AG)^T = AG; \\
&\text{iv)} && (GA)^T = GA
\end{aligned}
$$

则称矩阵G为矩阵A的加号逆或Moore-Penrose广义逆，记作A^+。如果G满足定义中的第一个条件，则称为A的减号逆，记作A^-。

3.1 广义逆的定义和性质

定理 (n, m)矩阵A的加号逆存在且唯一。

证明: 若A不是零矩阵（所有元素都等于零的矩阵），则A有奇异值分解$A = VDU^T$，取$G = UD^+V^T$，其中D^+是把对角阵D的主对角线中非零元素换成相应元素的倒数，其它元素保持为零，容易验证G是A的加号逆。当A为零矩阵时，$m \times n$的零矩阵是A的加号逆。

若A_1^+和A_2^+是$n \times m$矩阵A的两个加号逆，则

$$
\begin{aligned}
& A_1^+ = A_1^+ A A_1^+ = A_1^+ (A A_1^+)^T = A_1^+ (A_1^+)^T (A)^T = A_1^+ (A_1^+)^T (A A_2^+ A)^T \\
= & A_1^+ (A_1^+)^T A^T (A A_2^+)^T = A_1^+ (A A_1^+)^T A A_2^+ = A_1^+ A A_1^+ A A_2^+ = A_1^+ A A_2^+ \\
= & (A_1^+ A)^T A_2^+ = A^T (A_1^+)^T A_2^+ = (A A_2^+ A)^T (A_1^+)^T A_2^+ = (A_2^+ A)^T A^T (A_1^+)^T A_2^+ \\
= & (A_2^+ A)^T (A_1^+ A)^T A_2^+ = A_2^+ A A_1^+ A A_2^+ = A_2^+ A A_2^+ = A_2^+
\end{aligned}
$$

可见加号逆存在唯一。证毕。

加号逆有如下的性质:

i) $(A^+)^+ = A$;

ii) $(A^T)^+ = (A^+)^T$;

iii) $(\lambda A)^+ = \lambda^{-1} A^+,\ \forall \lambda \neq 0$;

iv) $\operatorname{rank}(A^+) = \operatorname{rank}(A) = \operatorname{rank}(AA^+) = \operatorname{rank}(A^+ A)$;

v) $(A^T A)^+ = A^+ (A^+)^T$;

vi) $A^+ = (A^T A)^+ A^T = A^T (AA^T)^+$;

vii) AA^+和$A^+ A$都是对称幂等矩阵;

viii) 若A是对称幂等矩阵，则 $A^+ = A$。

3.2 叉积阵

设A为(n, m)矩阵，

$A^T A$----线性模型的正规方程， 协方差阵估计

$A^T A$是对称半正定矩阵。$A^T A$与X之间有密切的联系。$\mathrm{rank}(A^T A) = \mathrm{rank}(A)$, $A = 0$当且仅当$A^T A = 0$，当$\mathrm{rank}(A) = p$时$A^T A$是对称正定阵。

若G是$A^T A$的一个减号逆，则G^T也是$A^T A$的减号逆（注意对称阵的减号逆不一定对称）：

$$(A^T A)G^T(A^T A) = [(A^T A)G(A^T A)]^T = (A^T A)^T = (A^T A)$$

$A^T A$的广义逆与A的广义逆密切相关。

对$A^T A$的任一个减号逆$(A^T A)^-$，$(A^T A)^- A^T$是A的减号逆。

3.3 正交投影

设$\mathbb{R}^n$为n维欧式空间，即$(x_1, x_2, \ldots, x_n), x_i \in \mathbb{R}, i = 1, 2, \ldots, n$这样的元素组成的集合并定义了加法和数乘，又定义了内积

$$(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T \boldsymbol{y} = \sum_{i=1}^{n} x_i y_i$$

当$(\boldsymbol{x}, \boldsymbol{y})$时称$\boldsymbol{x}$和$\boldsymbol{y}$正交，记作$\boldsymbol{x} \perp \boldsymbol{y}$。

设A为$n \times m$实数矩阵，记$A \in \mathbb{R}^{n\times m}$。称

$$\mu(A) = \{A\boldsymbol{\beta} \in \mathbb{R}^n : \boldsymbol{\beta} \in \mathbb{R}^m\}$$

为A的各列张成的线性子空间。$\mathbb{R}^n$的线性子空间是$\mathbb{R}^n$的子集且加法和数乘在此子集中封闭。记

$$\mu(A)^{\perp} = \{\boldsymbol{z} \in \mathbb{R}^n : A^T\boldsymbol{z} = \boldsymbol{0}\}$$

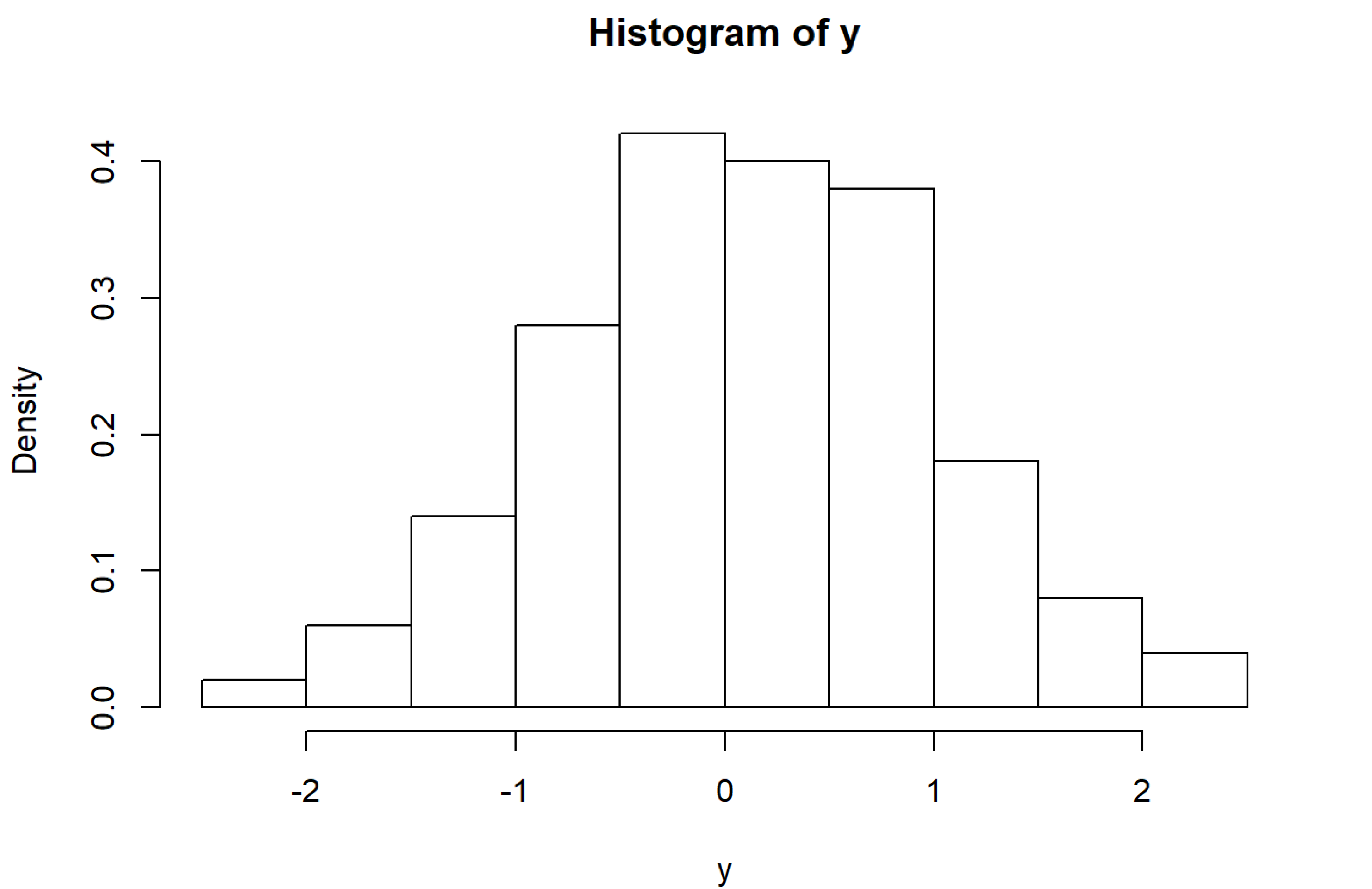
这是$\mathbb{R}^n$中所有与A的各列都正交的向量组成的集合，也是$\mathbb{R}^n$的子空间。

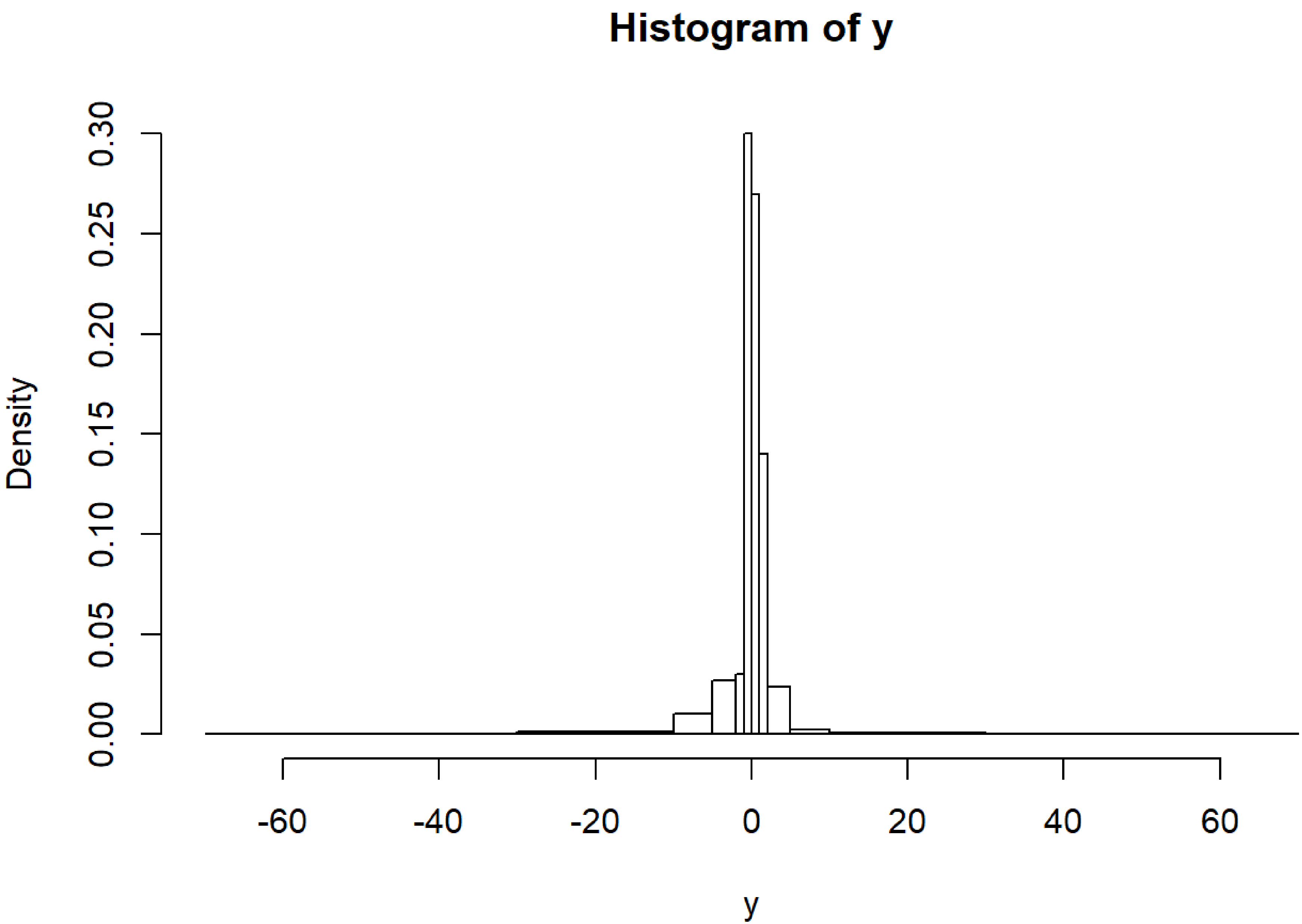
- 设P为n阶实对称矩阵，若P=P^2，则称**P为对称幂等阵**

若P是对称幂等阵，则P是$\mathbb{R}^n$向$\mu(P)$的正交投影阵

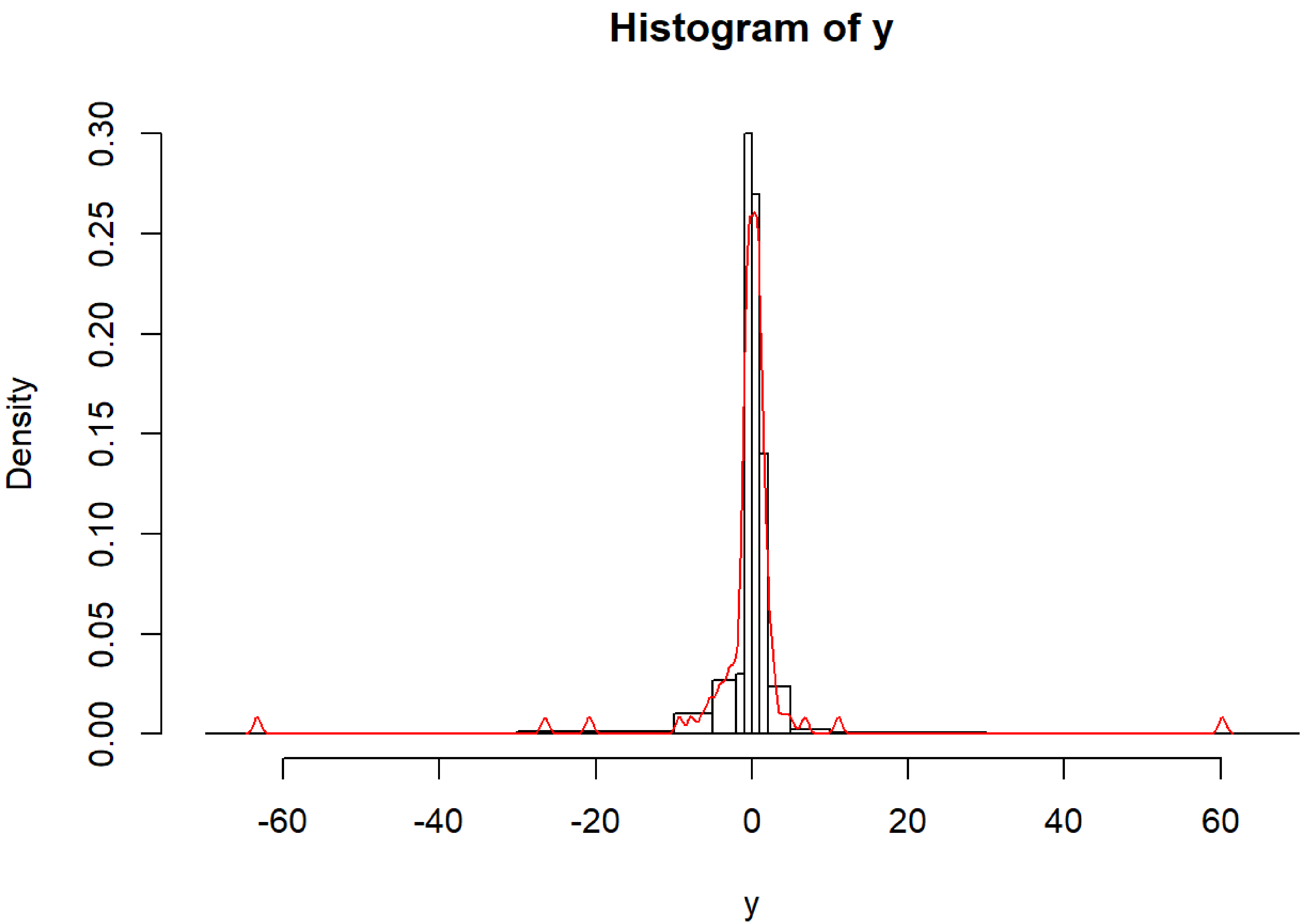
1.4 统计图形

1.4.1 直方图
1.4.2 核密度估计
1.4.3 盒形图
1.4.4 茎叶图
1.4.5 正态QQ图和正态概率图
1.4.6 散点图和曲线图
1.4.7散点图矩阵

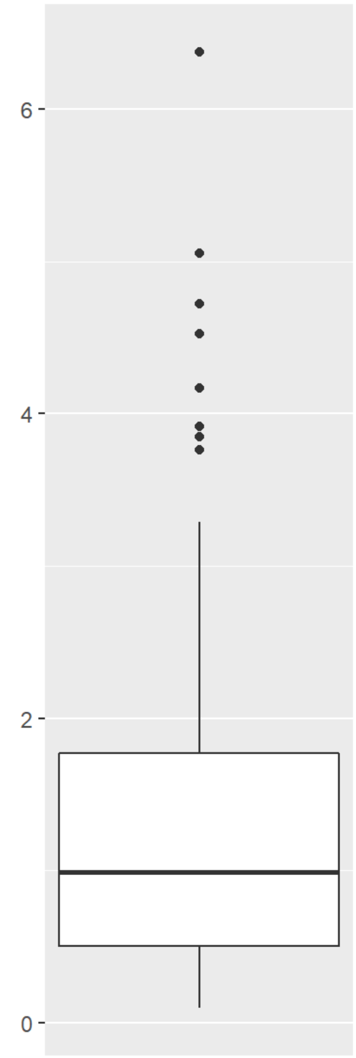

Histogram of y
Density
y
0.0
0.1
0.2
0.3
0.4
-2
-1
0
1
2



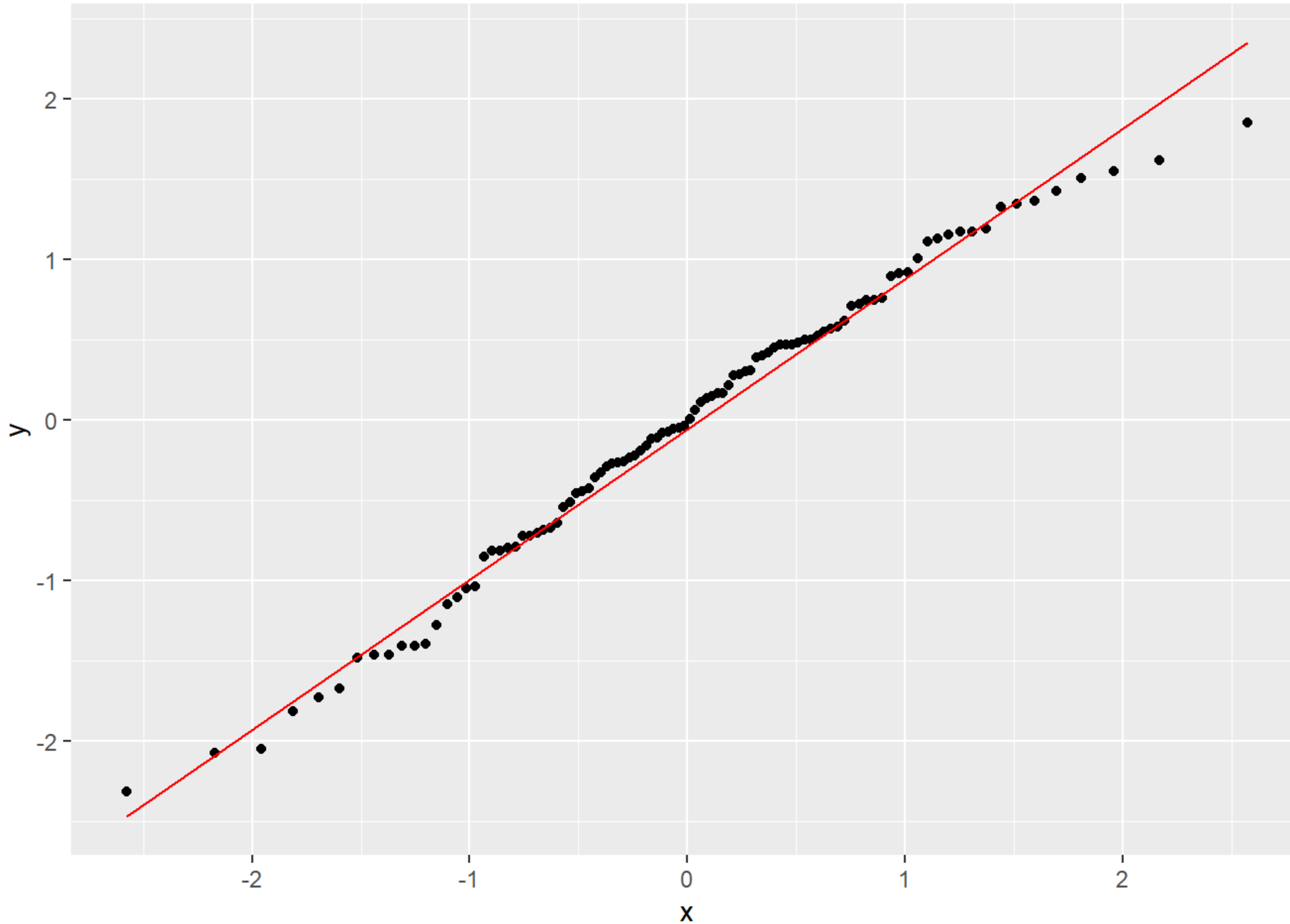
Histogram of y
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
-60
-40
-20
0
20
40
60
y


Histogram of y
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
-60
-40
-20
0
20
40
60
y

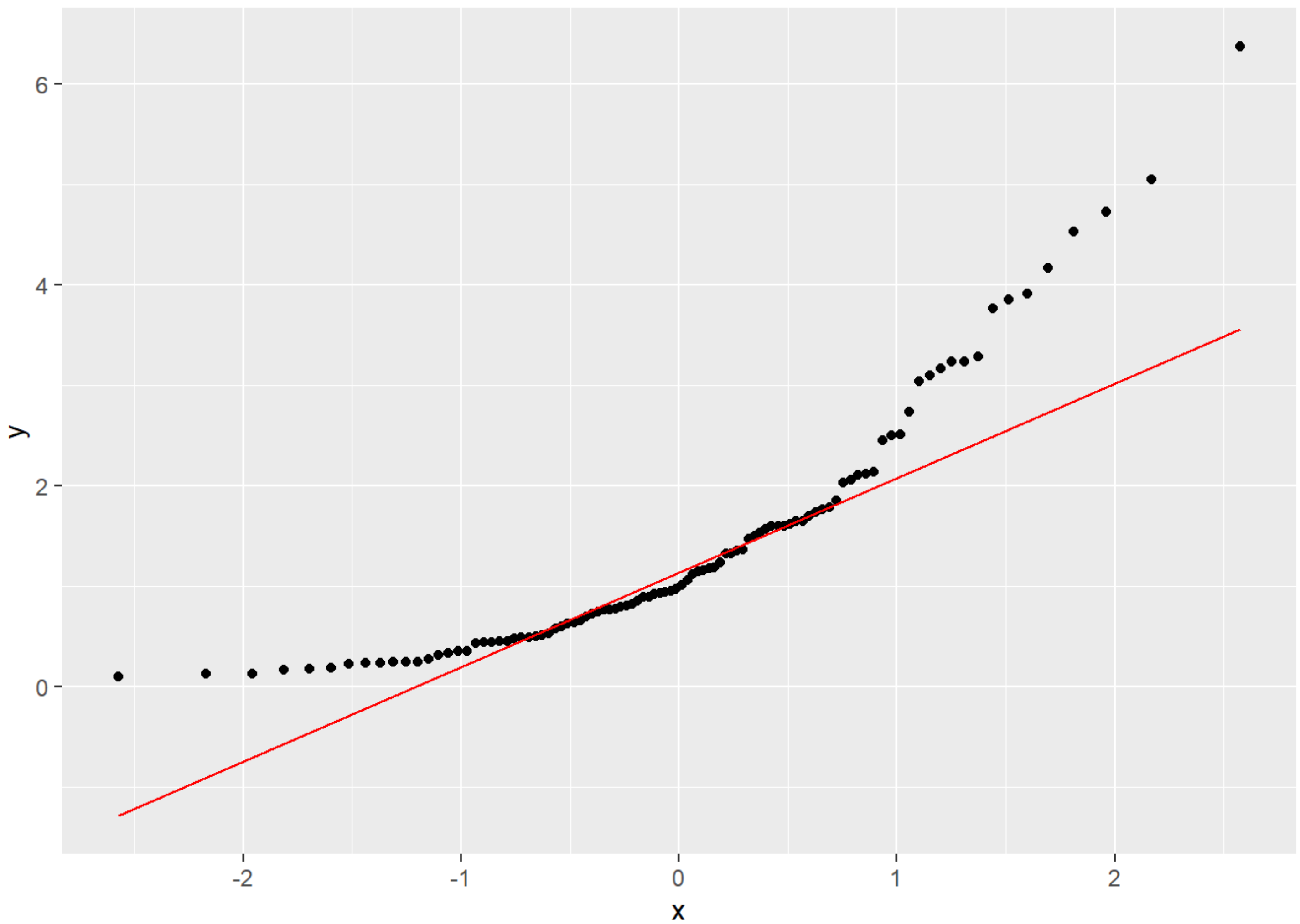
盒形图



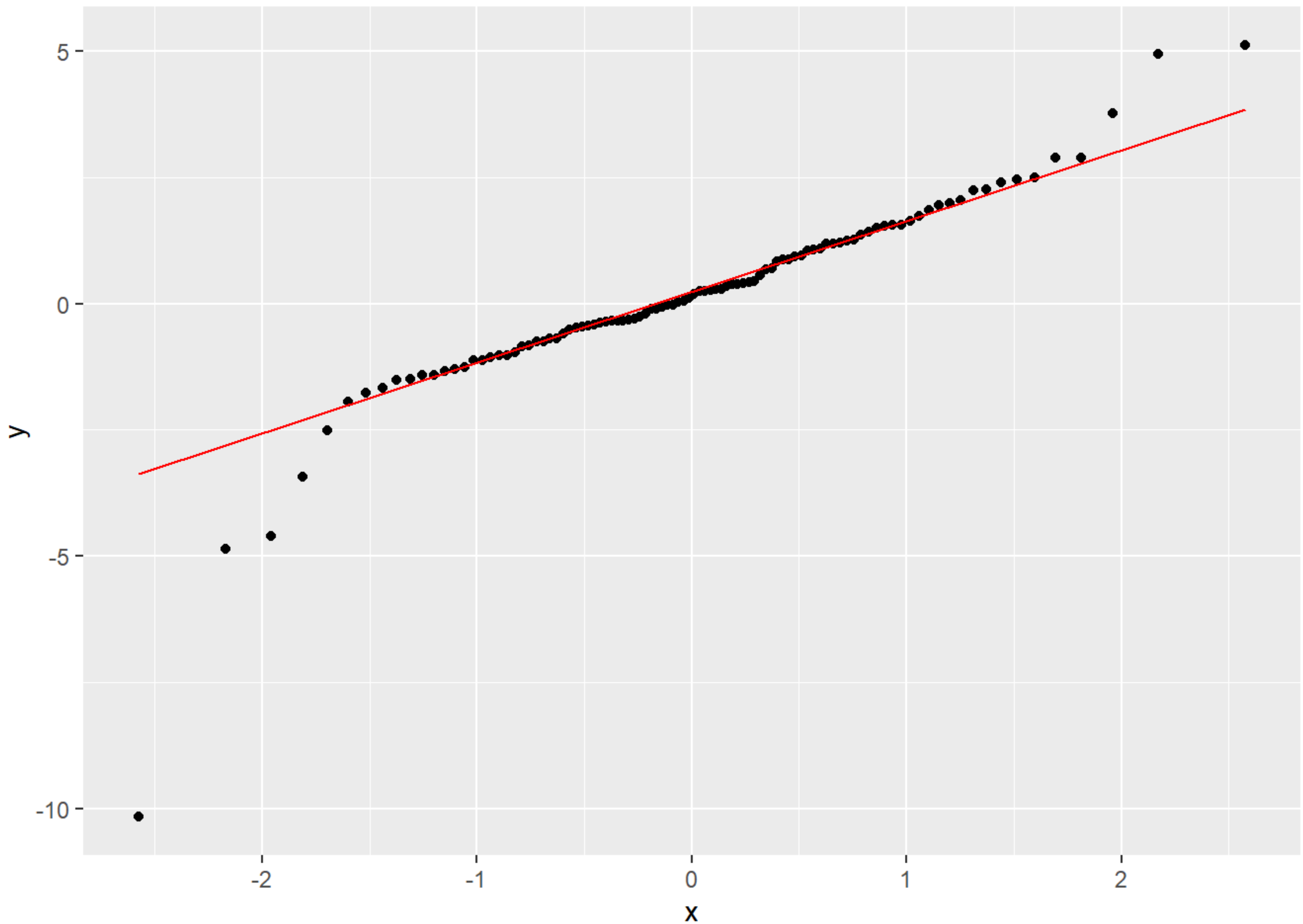
正态样本的正态QQ图



对数正态样本的正态QQ图



t分布样本的正态QQ图



三种不同的鸢尾花的150个样品的花瓣、花萼长、宽的数据的散点图矩阵: