统计学与 R 语言

第 19 讲回归分析 I

张敬信

2022 年 5 月 2 日

哈尔滨商业大学

回归分析（Regression Analysis），是统计学的核心算法，是计量模型和机器学习的最基本算法。

回归分析是确定两个或两个以上变量间相互依赖的定量关系的一种统计分析方法，具体是通过多组自变量和因变量的样本数据，拟合出最佳的函数关系。如果该关系是线性函数关系，就是线性回归。

计量模型和机器学习中的各种回归算法都可以看作是线性回归的扩展，分类算法也可以看作是一种特殊的回归。

回归分析常用于：

- 探索现象/结果的影响因素主要有哪些?
- 影响因素对现象/结果是怎样影响的?
- 预测未来的现象/结果

设 y 为因变量数据，$\mathbf{x}$ 为自变量数据（可以是多维），设二者之间的真实（精确）关系为：

$$y = f(\mathbf{x})$$

这是不可能得到的，所谓回归建模只是试图去找到一种近似的关系来代替它：

$$\hat{f}(\mathbf{x}) \approx f(\mathbf{x})$$

二者之差就是模型的残差：

$$\varepsilon = f(\mathbf{x}) - \hat{f}(\mathbf{x})$$

总是希望把 y 与 $\mathbf{x}$ 的关系都留在模型部分：$\widehat{f}(\mathbf{x})$，让残差部分最好只是白噪声（完全是随机误差，0 均值，微小标准差的正态分布）：

$$\varepsilon \sim N(0,\, \sigma_{\varepsilon}^2)$$

这说明建模成功；否则，就是模型尚未提取出充分的模型关系（欠拟合）。

构建的模型关系 $\widehat{f}(\mathbf{x})$, 可以是简单的线性关系（线性回归）、也可以是复杂的”黑箱” 模型（神经网络、支持向量机等），尽管无法得到精确的表达式，但仍可以用于预测。

回归建模的基本原则是：在没有显著差异的情况下，优先选择更简单的模型。简单模型已足够充分建模，非要用更复杂的模型则会适得其反（过拟合），会降低模型的泛化（预测）能力。

一. 一元线性回归

只对一个自变量与因变量之间的线性关系建模，其基本形式为：

$$y = \beta_0 + \beta_1 x$$

一元线性回归的全部模型预测值可表示为：

$$\begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}_{n\times 2} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}_{2\times 1}$$

记

$$\hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix}_{n\times 1}, \quad X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}_{n\times 2}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}_{2\times 1}$$

则矩阵形式表示为

$$\hat{Y} = X\beta$$

于是，让总的预测误差最小的” 最小二乘法” 优化问题就表示为

$$\arg\min_{\beta} \ J(\beta) = \|Y - \hat{Y}\|^2 = \|Y - X\beta\|^2$$

其中，$\|\cdot\|$ 为向量的范数（长度)。同样地，$J(\beta)$ 的极小值，在其一阶偏导值等于 0 处取到，按矩阵求导法则计算，可得 $2X^TX\beta - 2X^TY = 0.$

若 X 满秩，则 X^TX 可逆，从而

$$\beta = (X^TX)^{-1}X^TY$$

二. 多元线性回归

推广到多元线性回归模型，可对多个自变量与因变量之间的线性关系建模：

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

多元线性回归是找一个超平面，到各个散点的距离总和最小：

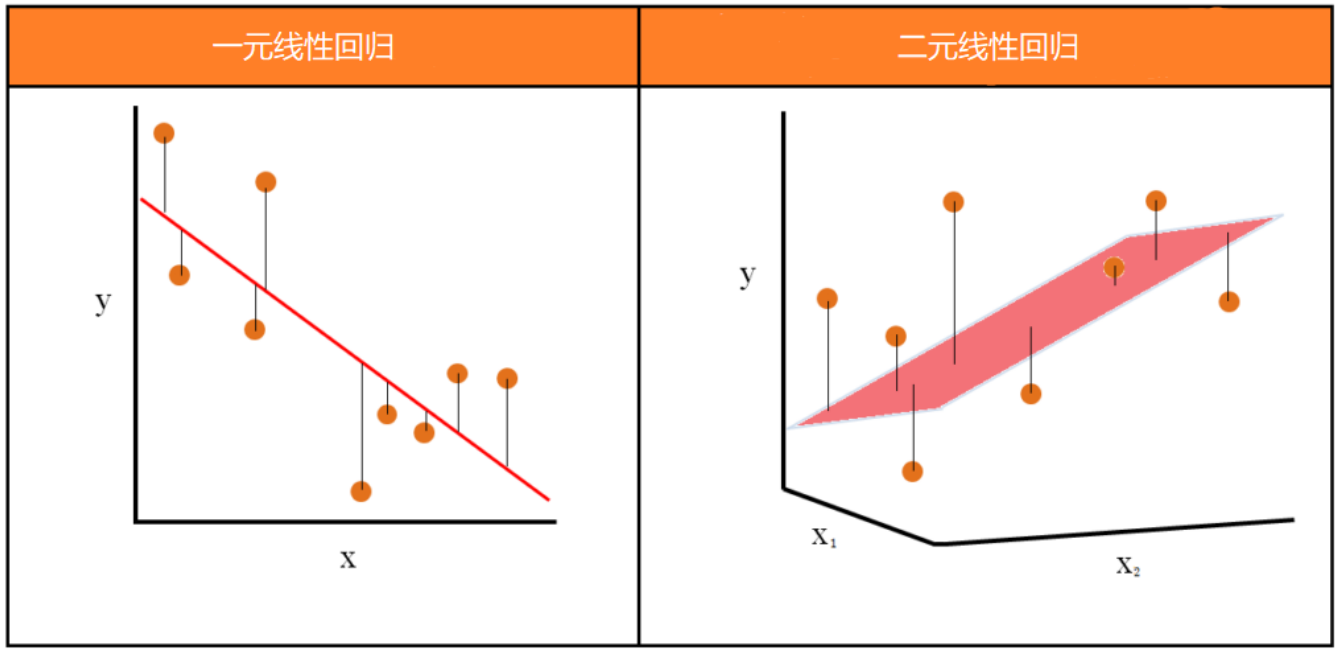


图 1: 线性回归示意图

m 个自变量，n 个样本，构成矩阵 X：

$$X = \begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_m^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(n)} & \cdots & x_m^{(n)} \end{bmatrix}$$

第 i 个样本为 $(x_1^{(i)}, \cdots, x_m^{(i)}, y_i)$. 令

$$\beta = (\beta_0, \beta_1, \cdots, \beta_m)^T$$

则 $\hat{Y} = X\beta$. 仍用最小二乘法找到最优的回归系数，结果形式不变：

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

称为**正规方程法**。

三. 回归诊断

线性回归模型的成功建模，依赖于如下的假设：

(1) 线性模型假设：$y = X\beta + \varepsilon$
(2) 随机抽样假设：每个样本被抽到的概率相同且同分布；
(3) 无完全共线性假设：X 满秩；
(4) 严格外生性假设：$E(\varepsilon \mid X) = 0$
(5) 球形扰动项假设：$Var(\varepsilon \mid X) = \sigma^2 I_n$
(6) 正态性假设：$\varepsilon \mid X \sim N(0, \sigma^2 I_n)$

其中，前三个是基础假设，严格外生性和球形扰动项假设分别保证了估计量的无偏性和有效性，正态性假设是为了进行统计推断做的额外假设:

- 前四个假设成立时，估计量无偏；
- 前五个假设成立时，估计量有效，是最优线性无偏估计量；
- 所有假设都成立时，估计量是最优估计量。

1. 拟合优度检验

计算 R^2, 也称为可决系数，反映了自变量所能解释的方差占总方差的百分比：

$$\mathrm{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \quad \mathrm{SSE} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2, \quad \mathrm{SSR} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$R^2 = \frac{\mathrm{SSE}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSR}}{\mathrm{SST}}$$

R^2 值越大说明模型拟合效果越好。

注： R^2 未考虑自由度问题，为避免增加自变量而高估 R^2，选择调整的 R^2 是更合理的：

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1} \cdot (1 - R^2)$$

其中，n 为样本数，p 为自变量个数。

2. 均方误差与均方根误差

均方误差：

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

均方根误差：

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

均方根误差，刻画的是预测值与真实值平均偏离多少，是所有回归模型（包括机器学习中的回归算法）最常用的性能评估指标。

3. 残差检验

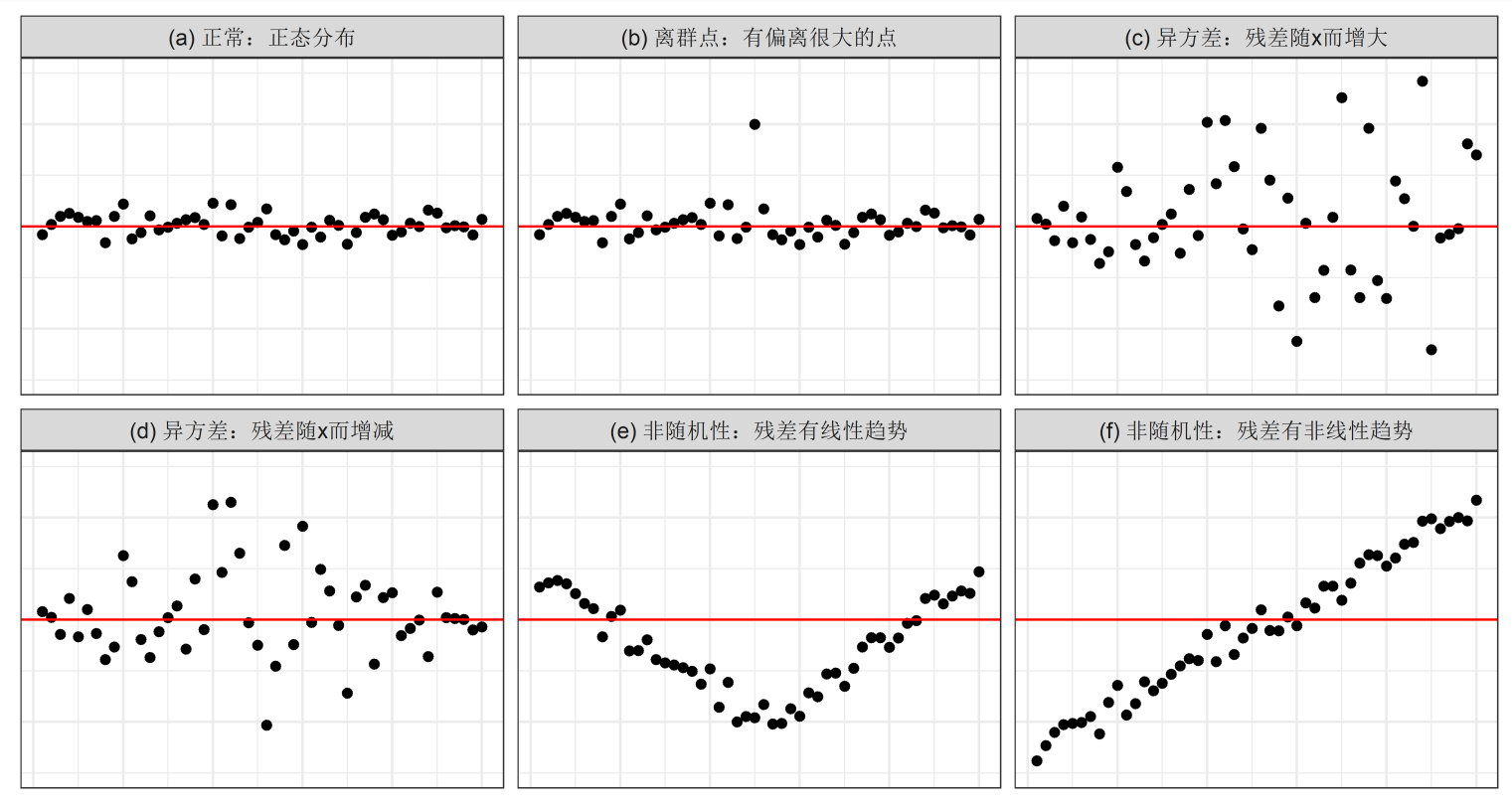


图 2: 残差分类图

- 只有图 (a) 说明模型是成功的，把模型部分都提取出来了；
- （e）和 (f) 属于模型有问题，没有把模型部分提取完全；
- （b）说明数据有异常点，应处理掉它重新建模；
- （c）残差随 x 的增大而增大，(d) 残差随 x 的增大而先增后减，都属于异方差。

(1) 残差正态性检验

用残差检验模型是否成功，就是对残差做正态性检验。也可以考察学生化残差(可回避标准化残差的方差齐性假设）是否服从标准正态分布。

(2) 残差独立性检验

残差是白噪声，也表明不具有自相关性 (独立性)。用 `Durbin-Watson` 检验：

$$H_0：\text{残差不存在自相关；} \quad H_1：\text{残差是相关的}$$

$$\mathrm{DW} = \sum_{i=2}^{n} \frac{(\varepsilon_i - \varepsilon_{i-1})^2}{\mathrm{ESS}}$$

用 `lmtest::dwtest()` 实现：

- $\mathrm{DW} \approx 0$，表示残差中存在正自相关；
- $\mathrm{DW} \approx 4$，表示残差中存在负自相关；
- $\mathrm{DW} \approx 2$，表示残差不存在自相关。

若残差存在自相关性，则需要考虑给模型增加自回归项。

(3) 异方差检验

线性回归的模型假设包括 $Var(\varepsilon \mid X) = \sigma^2 I_n$, 即要求残差的方差是不随样本而变化的相同值 σ^2，否则就称为残差具有异方差性。

检验残差的异方差性，可用 `Breusch-Pagan` 检验，原假设是不存在异方差。用 `lmtest::bptest()` 实现。

异方差将导致回归系数的标准误估计错误，一种解决办法是估计异方差–稳健标准误。另一种是在回归之前对数据 y 或 x 进行变换，实现方差稳定后再建模。原则上，当残差方差变化不太快时取开根号变换 $\sqrt{y}$；当残差方差变化较快时取对数变换 $\ln y$；当残差方差变化很快时取逆变换 $1/y$；还有其他变换，如著名的 `Box-Cox` 变换或 `Yeo-Johnson` 变换 (可应付负值)，将非正态分布数据变换为正态分布.

4. 共线性诊断

多元线性回归建模，若自变量数据之间存在较强的线性相关性，即存在**多重共线性**。

多重共线性，会导致回归模型不稳定，这样得到的回归模型，是伪回归模型，就是并不反映自变量与因变量的真实影响关系。

比如，真实模型关系是 $y = 2x_1 + 3x_2$, 若 x_1 与 x_2 存在线性关系：$x_2 = 2x_1$，则建模成 $y = 4x_1 + 2x_2$, $y = 6x_1 + x_2$, $y = 8x_1, \cdots$ 都完全没有问题。

多元线性回归建模，需要做**共线性诊断**，识别出多重共线性，并处理多重共线性再建模。

从线性相关系数、回归模型的方差膨胀因子 VIF（大于 10）来确定：

$$\text{VIF} = \frac{1}{1 - R_j^2}$$

它是 $\text{Var}(\hat{\beta}_j)$ 的决定性因子，其中 R_j 是第 j 个自变量与其余自变量之间的可决系数，R_j^2 越接近 1, 说明该变量越能被其余变量所解释。

多重共线性的解决办法（任选其一）：

- 若两个自变量线性相关系数较大，则只用其中一个自变量；
- 用逐步回归，剔除冗余的自变量，得到更稳健的回归模型；
- 用主成分回归，相当于对自变量进行重组（将线性相关性强的变量合成为主成分），再做线性回归；
- 利用正则化回归：岭回归、Lasso 回归、弹性网模型（岭回归与 Lasso 回归的组合）

(1) 回归系数的显著性

回归方程反映了因变量 y 随自变量 x 变化而变化的规律，若其系数 $\beta_1 = 0$，则 y 不随 x 变化，此时回归方程无意义。所以，要做 β_1 是否显著非 0 的假设检验：

$$H_0: \beta_1 = 0, \quad H_1: \beta_1 \neq 0$$

F 检验[1]：若 H_0 为真，则回归平方和 RSS 与残差平方和 $\frac{\mathrm{ESS}}{n-2}$ 都是 σ^2 的无偏估计，构造 F 统计量：

$$F = \frac{\mathrm{RSS}/\sigma^2/1}{\mathrm{ESS}/\sigma^2/(n-2)} = \frac{\mathrm{RSS}}{\mathrm{ESS}/(n-2)} \sim F(1, n-2)$$

来检验原假设 $\beta_1 = 0$ 是否为真。

[1]也可以用的 t 检验，与 F 检验是等价的，因为 $t^2 = F$.

(2) 回归标准误与回归系数标准误

回归模型的标准误，衡量的是以样本回归直线为中心分布的观测值同直线上拟合值的平均偏离程度：

$$s = \sqrt{\frac{\mathrm{SSE}}{n-p}} = \sqrt{\frac{\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon}_i)^2}{n-p}} = \sqrt{\frac{\sum_{i=1}^{n}\varepsilon_i^2}{n-p}}$$

其中，SSE 为残差平方和，n 为样本数，$n-p$ 为自由度，p 为包括常数项在内的自变量的个数。

回归系数标准误（抽样误差的标准差），是对回归系数这一估计量标准差的估计值，衡量的是在一定的样本量下，回归系数同其期望的平均偏离程度[2]：

$$SE(\hat{\beta}_k) = \sqrt{\mathrm{Var}(\hat{\beta}_k)} = \sqrt{s^2(X^TX)^{-1}_{kk}}$$

[2]该标准误是来自统计学家得到的理论公式，另一种方法是用 Bootstrap 法.

6. 回归模型预测

通过检验的回归模型，就可以用来做预测：将新的自变量数据代入回归模型计算 y

例如，得到一元线性回归方程 $\hat{y} = \beta_0 + \beta_1 x$ 后，预测 $x = x_0$ 处的 y 值为 $\hat{y}_0 = \beta_0 + \beta_1 x_0$，其置信区间[3]为：

$$(\hat{y}_0 - t_{\alpha/2}\sqrt{h_0\hat{\sigma}^2},\ \hat{y}_0 + t_{\alpha/2}\sqrt{h_0\hat{\sigma}^2})$$

其中，$t_{a/2}$ 的自由度为 $n-2$, $h_0 = \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$ 称为杠杆率，$\hat{\sigma}^2 = \frac{\mathrm{ESS}}{n-2}$.

[3]该置信区间是基于理论公式，也可以用 Bootsrap 法.

补充：相关分析

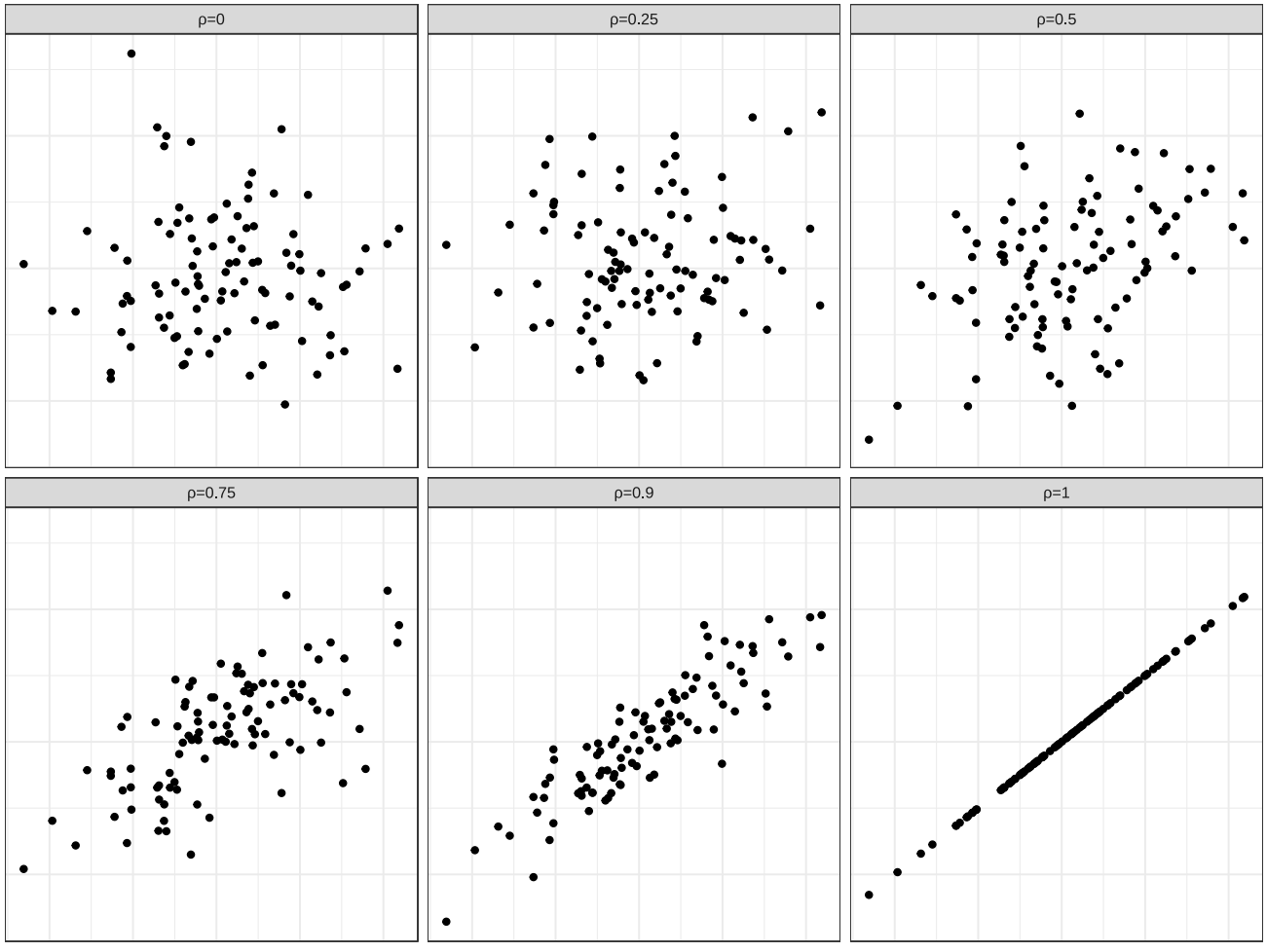
1. Pearson **线性相关系数**
   - **协方差**能反映两个正态连续变量的相互影响关系：

$$\mathrm{Cov}(X,Y)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x_i})(y_i-\bar{x_i})$$

   但是协方差的单位是不一致的，不具有可比性，解决办法就是做标准化，得到**相关系数**：

$$\rho_{XY}=\frac{\mathrm{Cov}(X,Y)}{s_X s_Y}=\frac{\sum_{i=1}^{n}(x_i-\bar{x_i})(y_i-\bar{x_i})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

线性相关系数介于 -1 和 1 之间，反映了线性相关程度的大小。


ρ=0
ρ=0.25
ρ=0.5
ρ=0.75
ρ=0.9
ρ=1

用 rstatix 包计算相关系数矩阵，并去掉重复，按相关系数大小排序：

```
library(rstatix)
cor_mat(iris[-5]) %>%                    # 相关系数矩阵
  replace_triangle(by = NA) %>%          # 将下三角替换为 NA
  cor_gather() %>%                       # 宽变长
  arrange(- abs(cor))                    # 按绝对值降序排列
#>           var1          var2   cor        p
#> 1 Petal.Length  Petal.Width  0.96 4.68e-86
#> 2 Sepal.Length Petal.Length  0.87 1.04e-47
#> 3 Sepal.Length  Petal.Width  0.82 2.33e-37
#> 4  Sepal.Width Petal.Length -0.43 4.51e-08
#> 5  Sepal.Width  Petal.Width -0.37 4.07e-06
#> 6 Sepal.Length  Sepal.Width -0.12 1.52e-01
```

注意：统计相关并不代表因果相关！线性不相关也可能具有非线性关系！

2. Spearman **秩相关系数**

不符合正态分布，只能考虑对数据小到大编秩，计算 Spearman 秩相关系数：

3. Kendall **秩相关系数**

对于分类变量，可以计算 Kendall 秩相关系数。

`rstatix::cor_test()` 函数，设置参数 `method` 为"spearman", "kendall" 即可实现相应相关系数计算。

4. **其它相关**：

- **偏相关**：控制一些变量之后，考察两变量的相关性
- **典型相关**：两组（线性组合）变量之间的相关性

另外，`correlationfunnel` 包能够快速探索自变量，特别是大量分类变量，对因变量的相关性影响大小，并绘制”相关漏斗图”进行可视化。

本篇主要参阅 (张敬信, 2022), 以及包文档，模板感谢 (黄湘云, 2021), (谢益辉, 2021).

参考文献

张敬信 (2022). *R* 语言编程：基于 *tidyverse*. 人民邮电出版社, 北京.

谢益辉 (2021). *rmarkdown: Dynamic Documents for R*.

黄湘云 (2021). *Github: R-Markdown-Template*.