北京航空航天大学
经济管理学院
BEIHANG UNIVERSITY
SCHOOL OF ECONOMICS AND MANAGEMENT

第 2.1 讲：相关系数

康雁飞

数量经济与商务统计系

Outline

相关关系

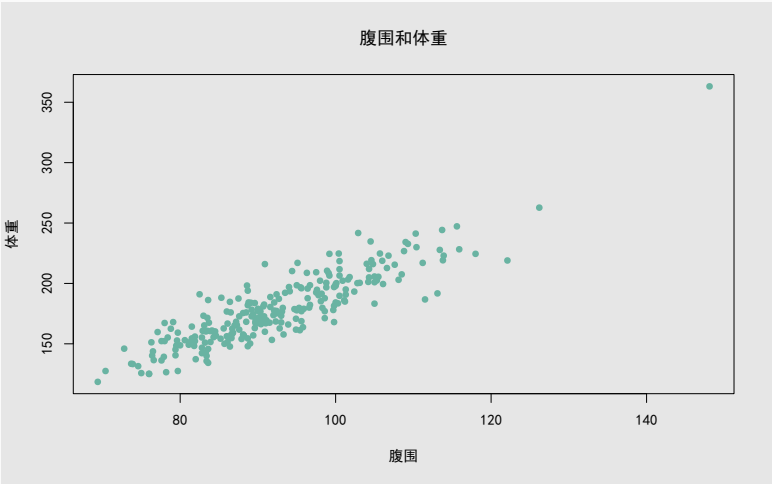
函数关系

每一个 X 值都唯一地对应一个 Y 值，即 $Y = f(X)$。

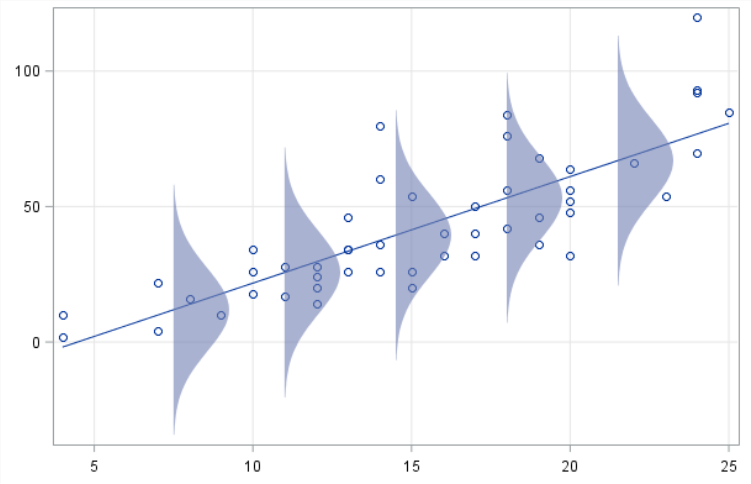
随机关系

给定 X 的条件下，Y 的取值服从一个概率分布。

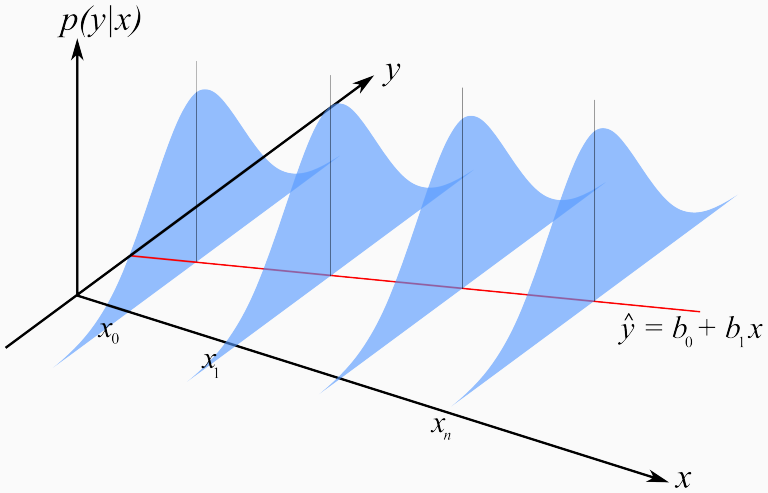
腹围与体重



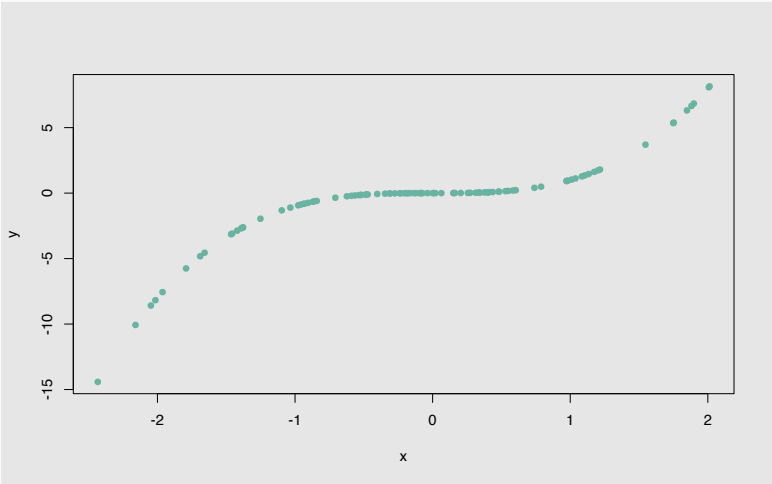
随机关系



随机关系



非线性关系



Outline

Pearson 相关系数

Pearson 相关系数

给定观测值 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$，Pearson 相关系数定义为：

$$r_{xy} = \frac{s_{xy}}{s_x s_y},$$

其中 $s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$，$s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$，

$s_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2}$.

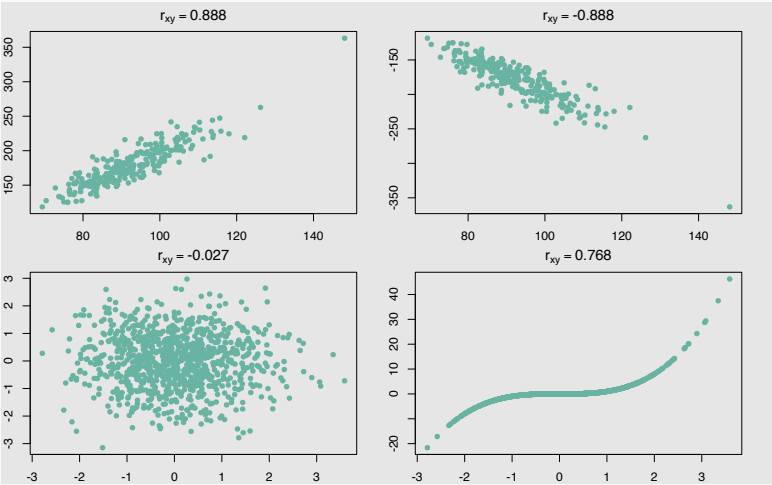
$-1 \leq r_{xy} \leq 1$

r	相关性
$r > 0$	正线性相关
$r < 0$	负线性相关
$r = 0$	线性无关
$r = 1$	完全正线性相关
$r = -1$	完全负线性相关

r_{xy} 的经验判断

r	相关性
$\lvert r_{xy}\rvert \geq 0.8$	强线性相关
$0.5 \leq \lvert r_{xy}\rvert < 0.8$	中度线性相关
$0.3 \leq \lvert r_{xy}\rvert < 0.5$	弱线性相关
$\lvert r_{xy}\rvert < 0.3$	不相关

r_{xy} 主要用于衡量变量之间的线性关系



总体相关系数

总体相关系数

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y},$$

其中 $\sigma_{xy} = E[(X - \mu_X)(Y - \mu_Y)]$, $\sigma_x^2 = E[(X - EX)^2]$, $\sigma_y^2 = E[(Y - EY)^2]$。

总体相关系数的假设检验

- $H_0: \rho = 0$
- $H_1: \rho \neq 0$

检验统计量为:

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

在 H_0 下，$T \sim t(n-2)$.

问题：如何做总体相关系数的假设检验?

R 中计算 Pearson 相关系数

```
cor(fat$Abdomen, fat$Weight)
```

```
## [1] 0.8879949
```

R 中进行总体相关系数检验

```
cor.test(fat$Abdomen, fat$Weight)
```

```
##
##  Pearson's product-moment correlation
##
## data:  fat$Abdomen and fat$Weight
## t = 30.532, df = 250, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal t
## 95 percent confidence interval:
##  0.8586427 0.9115424
```

Outline

Spearman 秩相关系数

- **Pearson** 相关系数只能测量两个随机变量之间是否存在线性相关关系！
- 如何测量非线性相关关系？

Spearman 相关系数

Spearman 相关系数

给定观测值 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$，Spearman 相关系数 ρ 定义为观测值的秩（Rank）之间的 Pearson 相关系数：

$$\rho_{xy} = \frac{\sum(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum(u_i - \bar{u})^2}\sqrt{\sum(v_i - \bar{v})^2}},$$

其中，u_i 和 v_i 代表 x_i 和 y_i 的秩。

例子

i	1	2	3	4	5	6
x_i	1	2	3	4	5	6
y_i	1	4	9	16	25	36

问题：**Pearson 相关系数 = ?**

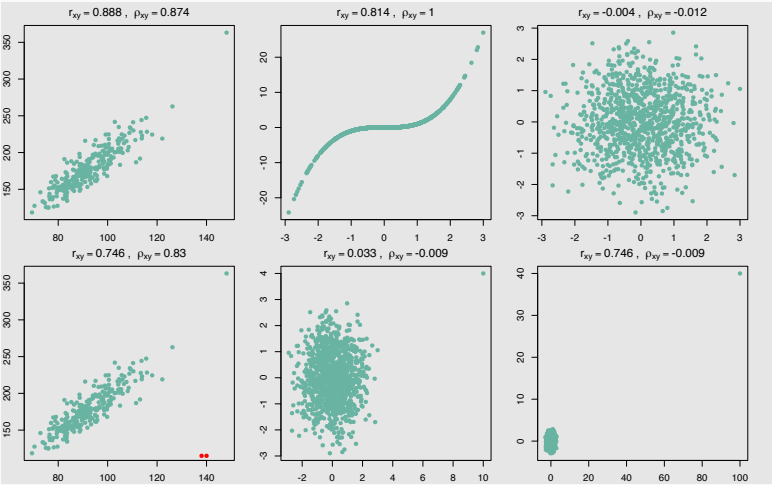
Spearman 相关系数

首先计算秩：

i	1	2	3	4	5	6
$\text{rank}(x_i)$	1	2	3	4	5	6
$\text{rank}(y_i)$	1	2	3	4	5	6

问题：**Spearman 相关系数 = ？**

r_{xy} 和 ρ_{xy}


rxy = 0.888 , ρxy = 0.874
rxy = 0.814 , ρxy = 1
rxy = -0.004 , ρxy = -0.012
rxy = 0.746 , ρxy = 0.83
rxy = 0.033 , ρxy = -0.009
rxy = 0.746 , ρxy = -0.009

总体相关系数的假设检验

检验统计量为：

$$T = \rho\sqrt{\frac{n-2}{1-\rho^2}}.$$

在 H_0 下，$T \sim t(n-2)$.

R 中计算 Spearman 相关系数

```
cor(fat$Abdomen, fat$Weight,
    method = 'spearman')
```

```
## [1] 0.8739719
```

R 中进行总体相关系数检验

```
cor.test(fat$Abdomen, fat$Weight,
         method = 'spearman')
```

```
##
##  Spearman's rank correlation rho
##
## data:  fat$Abdomen and fat$Weight
## S = 336133, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
```

Outline

Kendall τ 相关系数

Kendall τ 相关系数

给定观测值 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$，对于任意的 $i < j$，如果 $(x_i - x_j)(y_i - y_j) > 0$，$(x_i, y_i)$ 和 (x_j, y_j) 叫作同向数对。全部数据所有可能的数对为 $C_n^2 = \frac{n(n-1)}{2}$，其中同向数对记为 N_c，反向数对记为 N_d，Kendall τ 相关系数相关系数定义为：

$$\tau = \frac{N_c - N_d}{n(n-1)/2}.$$

R 中计算 Kendall τ 相关系数

```
cor(fat$Abdomen, fat$Weight,
    method = 'kendall')
```

```
## [1] 0.693417
```

R 中进行总体相关系数检验

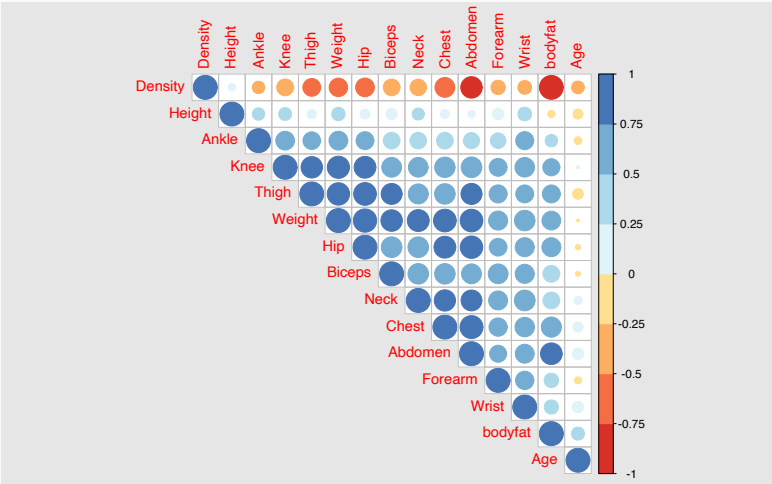
```
cor.test(fat$Abdomen, fat$Weight,
         method = 'kendall')
```

```
##
##  Kendall's rank correlation tau
##
## data:  fat$Abdomen and fat$Weight
## z = 16.359, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##      tau
```

相关系数在 R 中的可视化

```
library(corrplot)
library(RColorBrewer)
M <-cor(fat)
corrplot(M, type="upper", order="hclust",
         col=brewer.pal(n=8, name="RdYlBu"))
```

相关系数图矩阵


Density
Height
Ankle
Knee
Thigh
Weight
Hip
Biceps
Neck
Chest
Abdomen
Forearm
Wrist
bodyfat
Age
1
0.75
0.5
0.25
0
-0.25
-0.5
-0.75
-1

Outline

Pearson 相关系数

1. 衡量的是两个变量之间的线性相关关系
2. 一个介于-1 和 1 之间的取值，反映了变量之间的线性相关程度。正值表示正相关；负值表示负相关
3. 一般假设数据是连续数据，总体服从正态
4. 对离群值敏感

Spearman 相关系数

1. Spearman 相关系数适用于任意类型的可排序数据
2. 衡量的是两个变量之间的单调关系，而不是必须是线性关系
3. 计算时基于变量的秩次（秩次是排名的顺序），对原始数据进行秩次转换后再计算相关系数
4. 不受异常值影响：相对于 Pearson 相关系数，对离群值的影响更小

Kendall 相关系数

1. Kendall Tau 系数同样适用于任意类型的可排序数据
2. 也是衡量两个变量之间的单调关系，但与 Spearman 相关系数的计算方法略有不同
3. 计算时，基于比较变量对之间的相对大小关系来计算相关系数
4. 鲁棒性：对异常值的影响相对较小，更稳健。

Outline

以案例“智商与情商有关系吗？”（数据 iqeq.txt）比较几种相关系数及其相关性检验。