统计学原理(Statistic)

胡华平

西北农林科技大学

经济管理学院数量经济教研室

huhuaping01@hotmail.com

2021-05-16

第五章 相关和回归分析

5.1 变量间关系的度量

5.2 回归分析的基本思想

5.3 OLS方法与参数估计

5.4 假设检验

5.5 拟合优度与残差分析

5.6 回归预测分析

5.7 回归报告解读

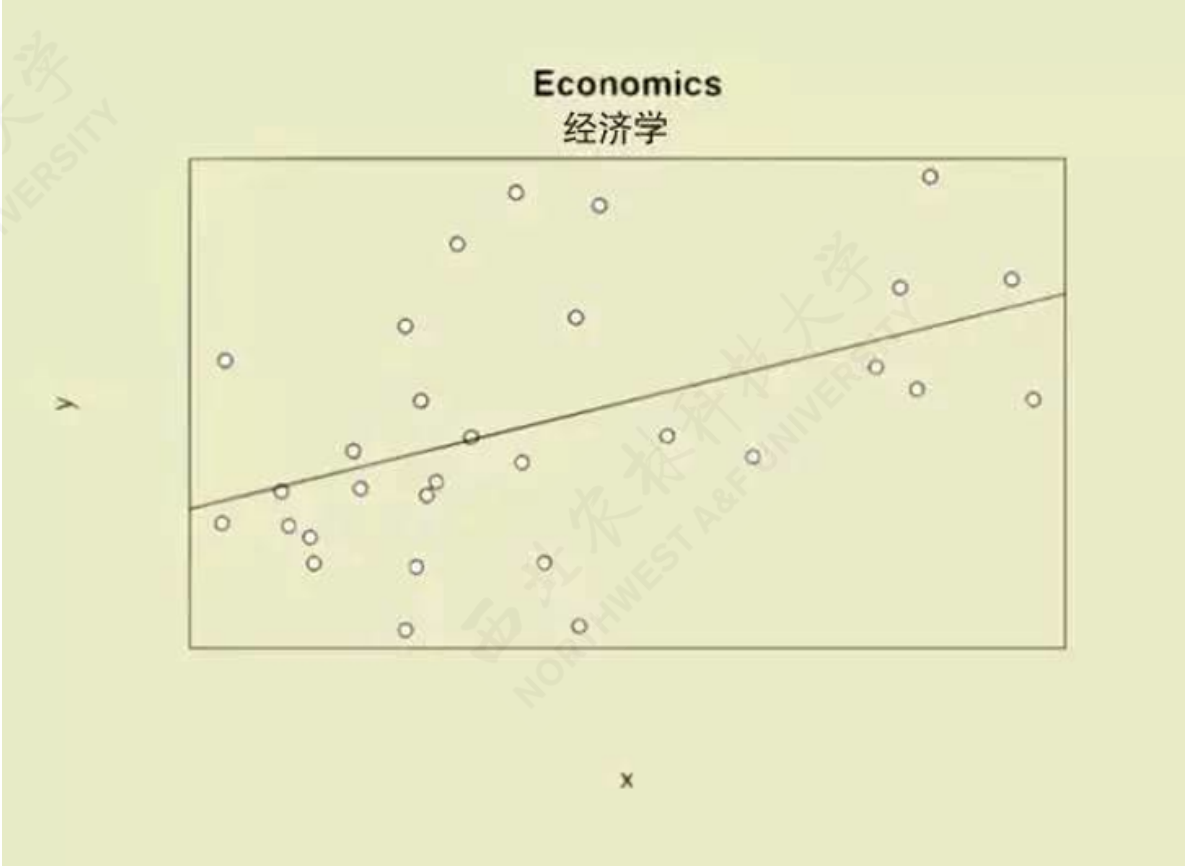
5.1 变量间关系的度量

变量间的关系

相关关系的描述与测度

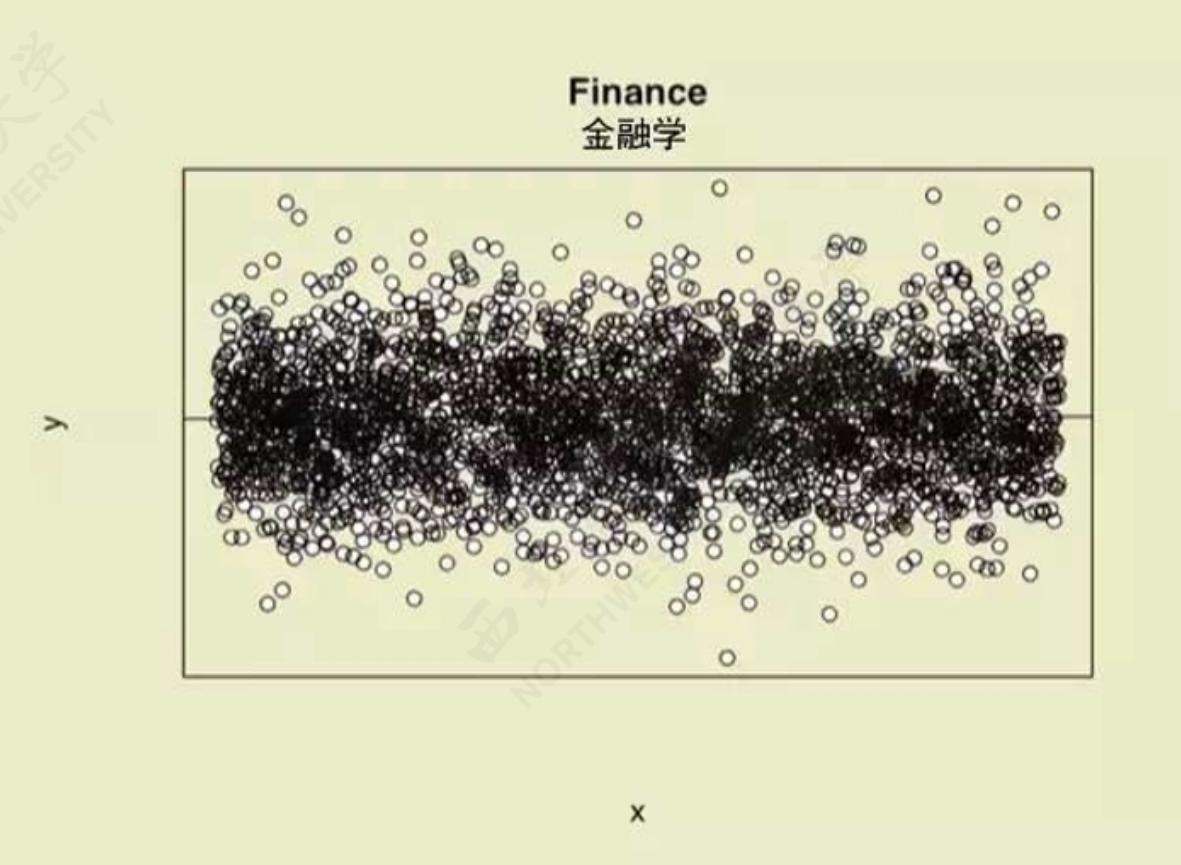
相关系数的显著性检验

（示例）变量间的关系：经济学专业解读



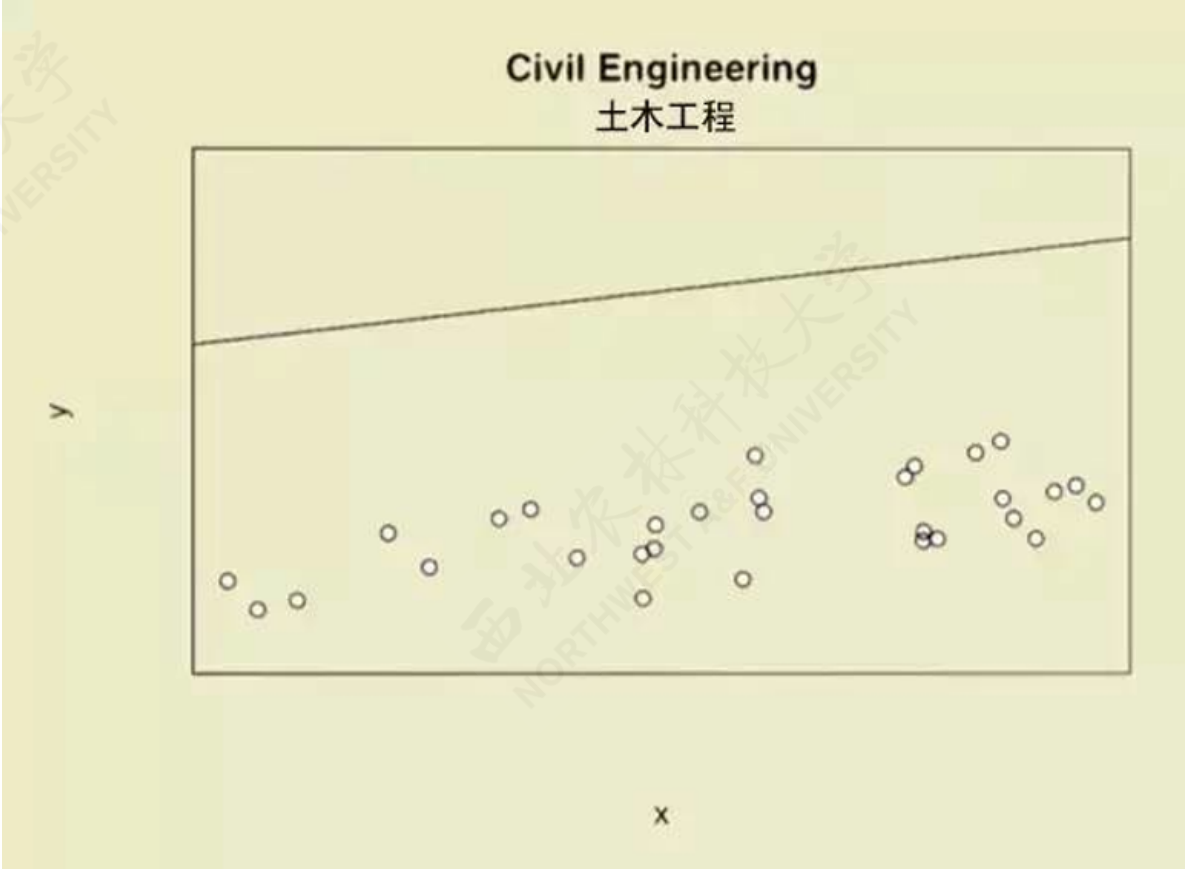
> “我们数据不少，做了很严格的回归，但异常值略多略多，符合理论的数值反而难找……”

（示例）变量间的关系：金融学专业解读



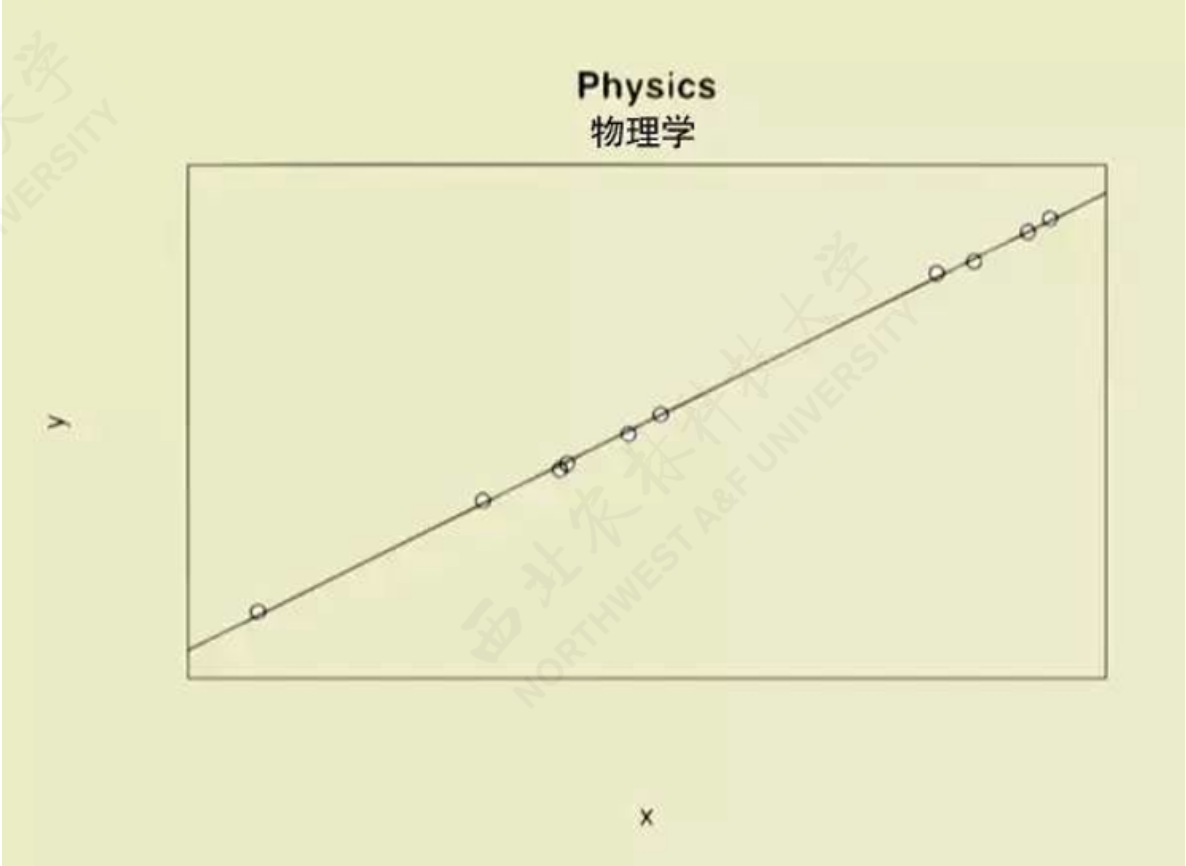
> “我们的数据多如牛毛，无孔不入。即使做完回归，也会发现异常值和符合理论的数值多得不忍直视。”

（示例）变量间的关系：土木工程专业解读



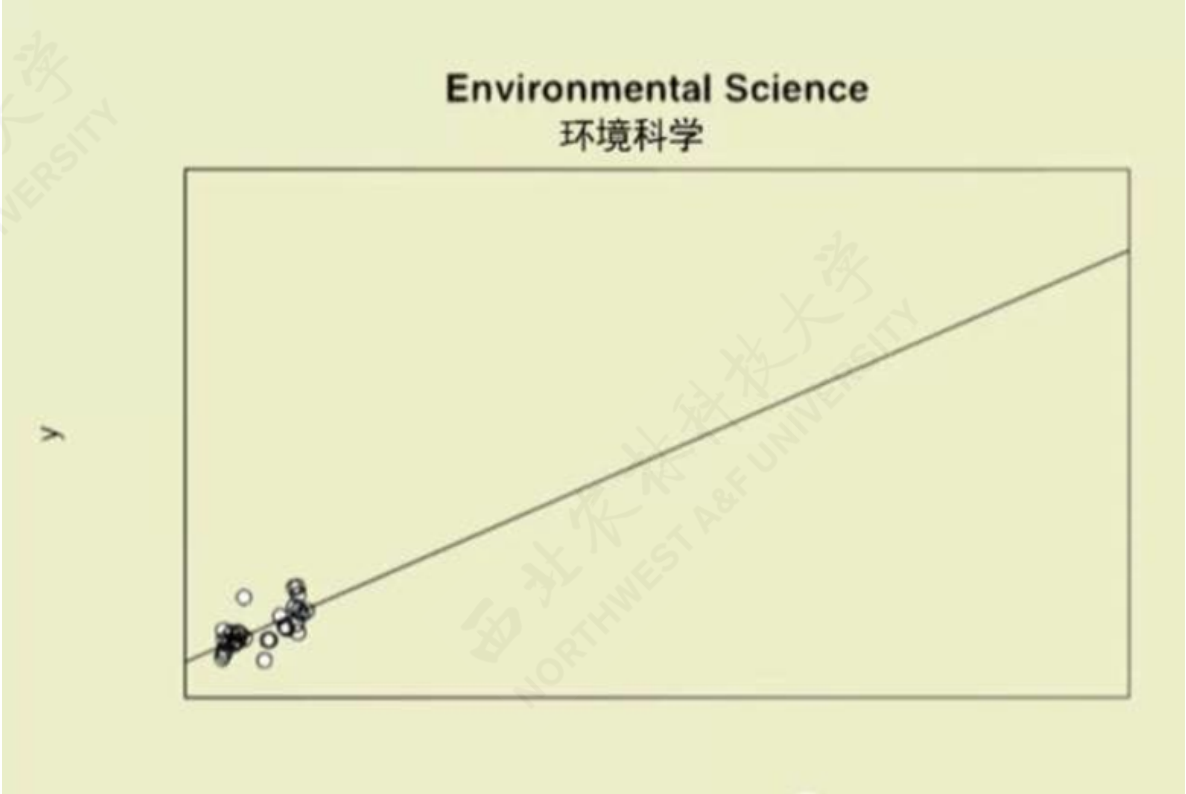
> “我们得要设计余量，所以理论设计得远高于实际承受……”

（示例）变量间的关系：物理学专业解读



> “我们的理论和数据严丝合缝，bingo！”

（示例）变量间的关系：环境科学专业解读

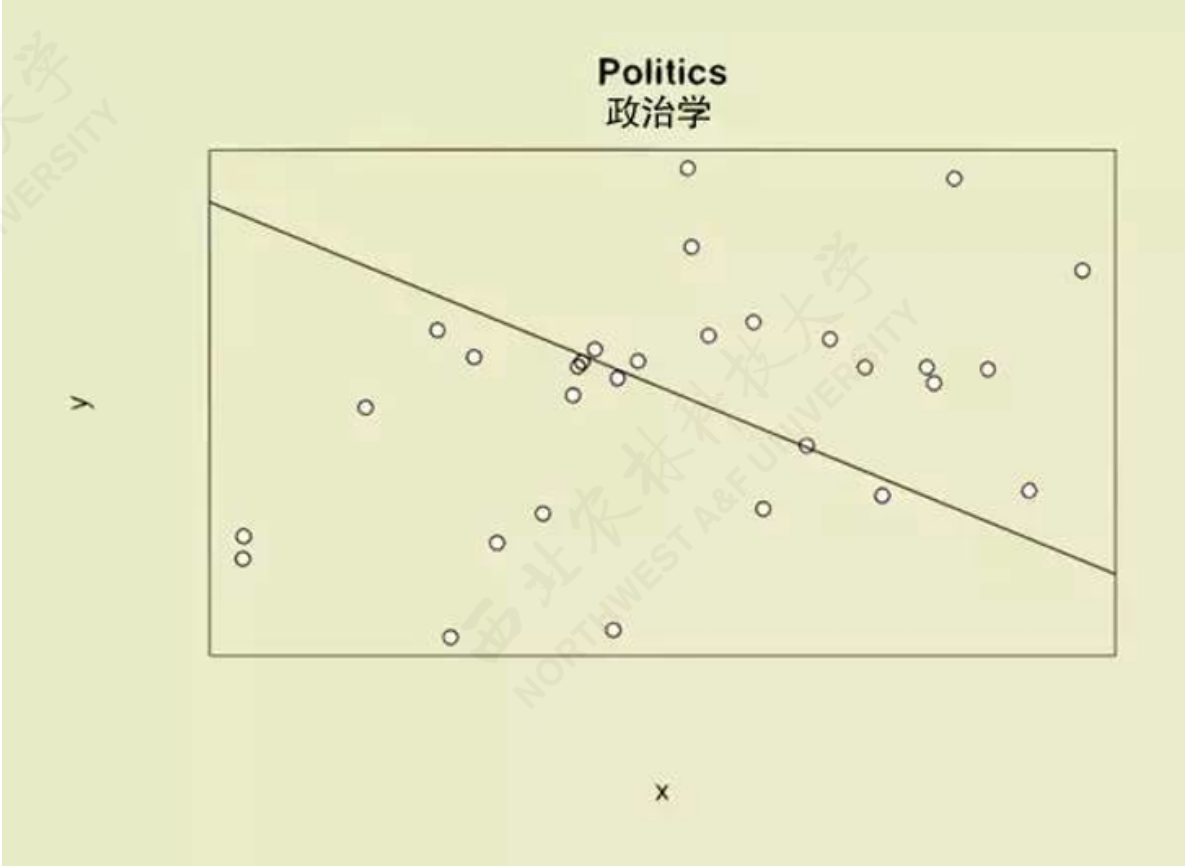


> “我们的理论和数据大致吻合，就是……应用范围有点蛋疼。”

（示例）变量间的关系：历史学专业解读

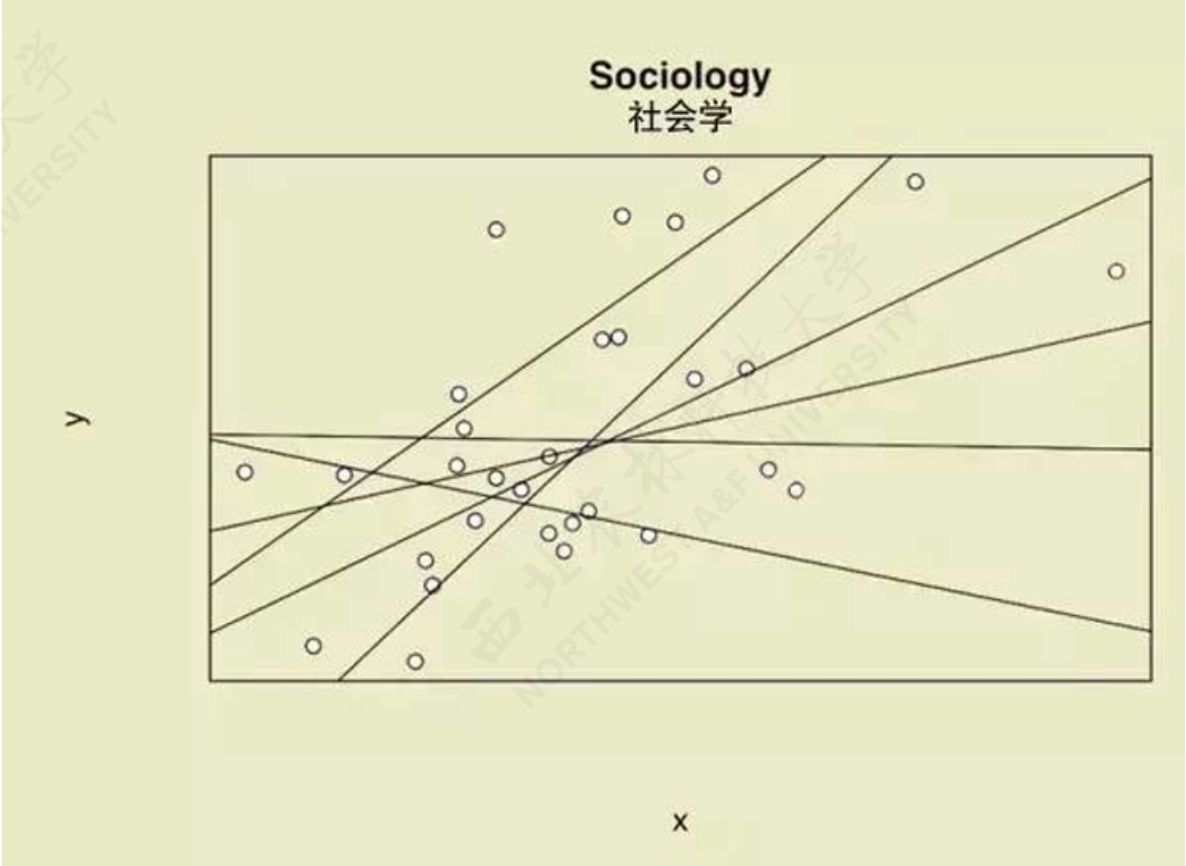
> “数据虽然很多，可我们能用理论把他们统统连起来！”

（示例）变量间的关系：政治学专业解读



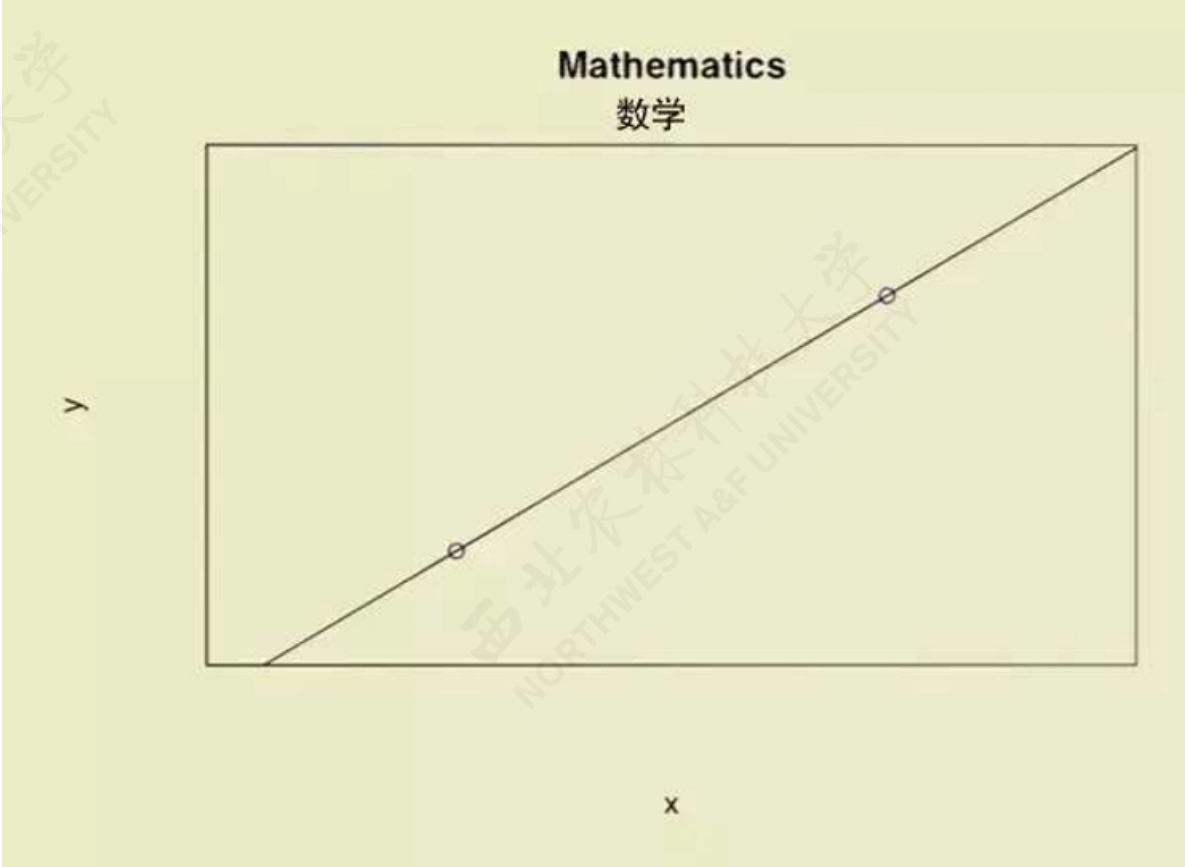
> “世界大势一日三变，尽管我们数据不少，可……我们的理论跟数据趋势是反着来的……”

（示例）变量间的关系：社会学专业解读



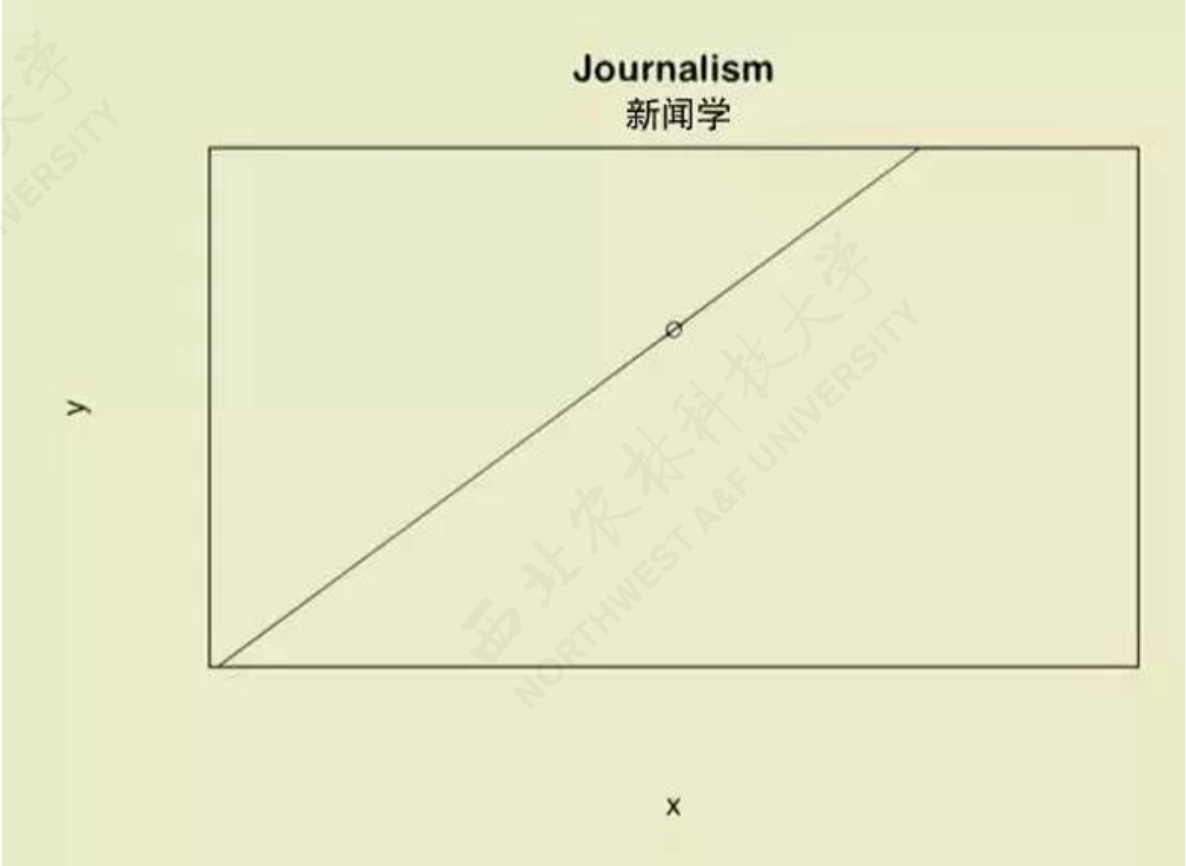
> “学海无涯苦作舟。那么多数据，那么多理论，慢慢学，恩……”

（示例）变量间的关系：数学专业解读



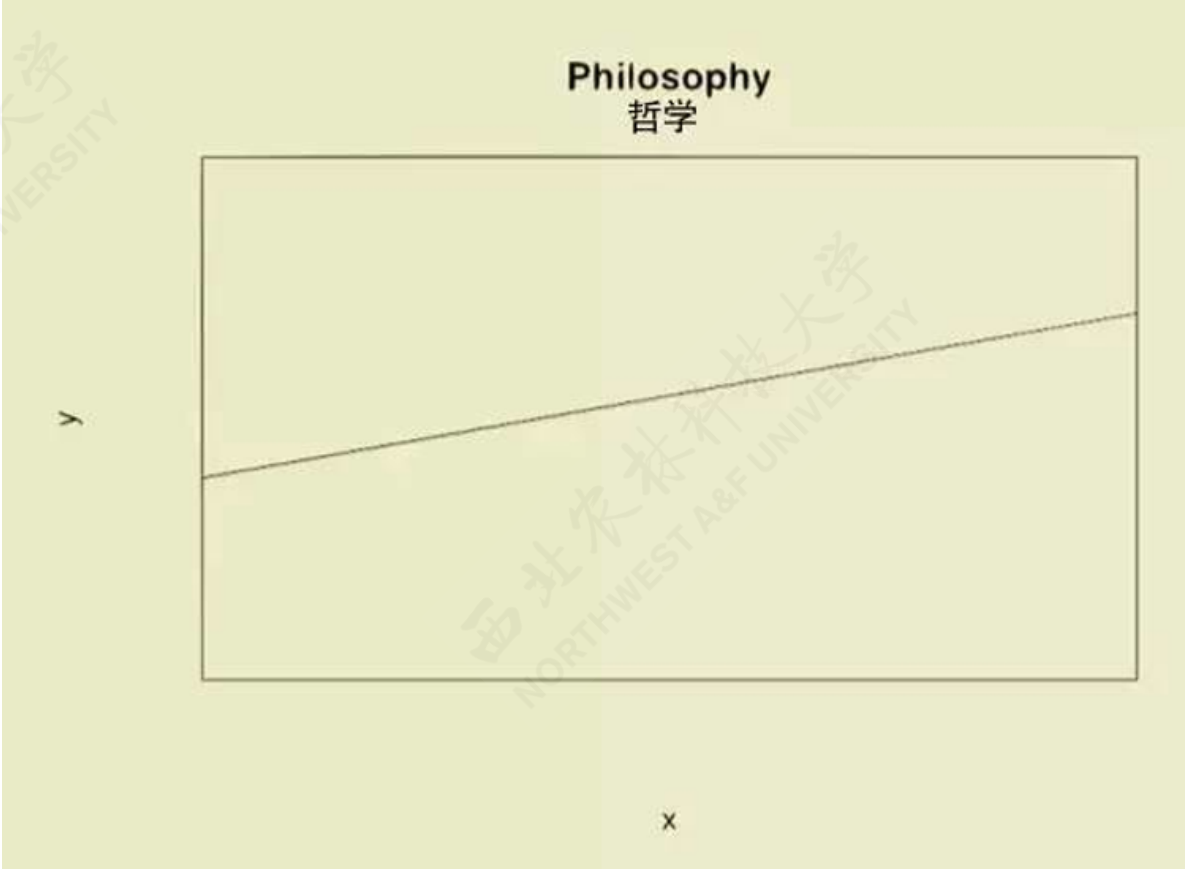
> “数据很少，但能建立理论～”

（示例）变量间的关系：新闻学专业解读



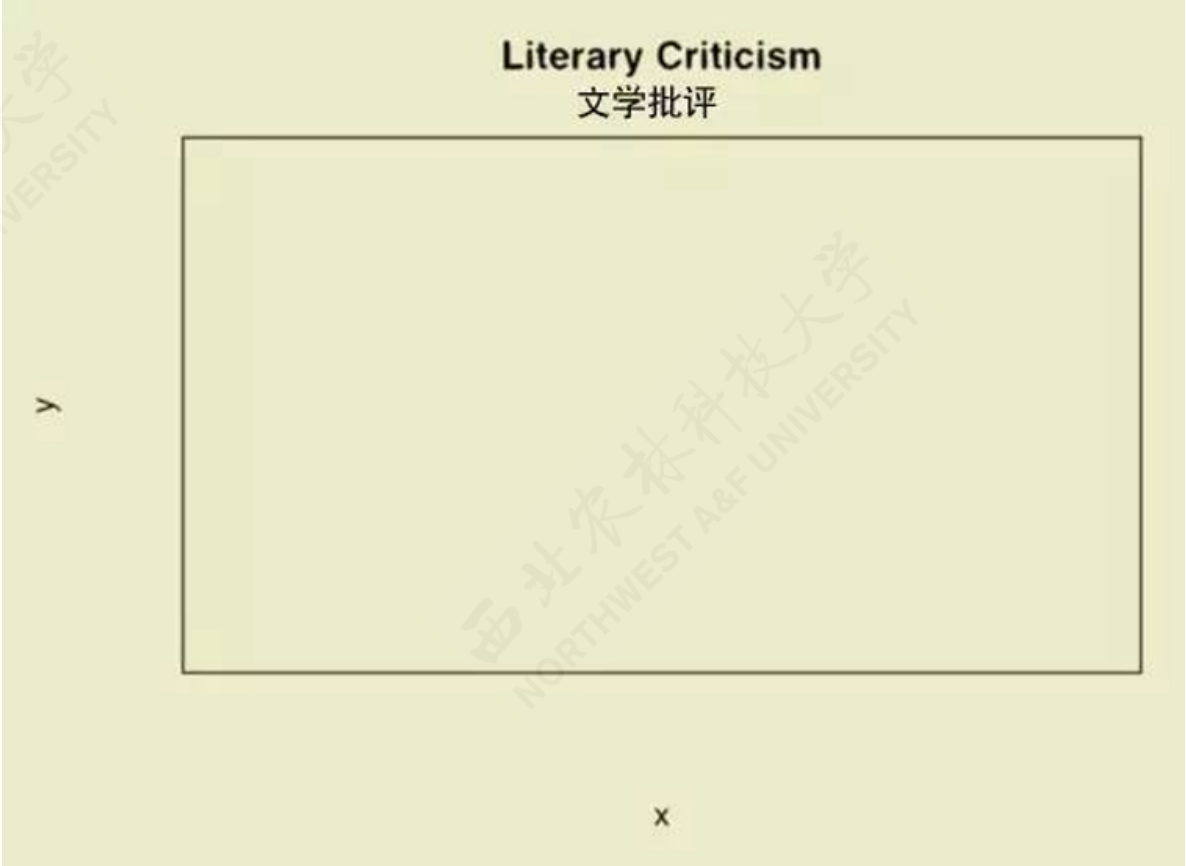
（示例）“只有一个数据，也能建立理论……”

（示例）变量间的关系：哲学专业解读



> “没有数据，依然建立理论……”

（示例）变量间的关系：文学批评专业解读



> “如图所示，你懂的……”

变量间的关系：函数关系

两个变量若存在是一一对应的确定关系，则称之为二者具有函数关系。

设有两个变量 X和 Y，变量 Y随变量 X一起变化，并完全依赖于 X，当变量 X取某个数值时， Y依确定的关系取相应的值，则称 Y是 X的函数，记为 $Y=f(X)$，其中 X称为自变量， Y称为因变量。

> 从几何学角度来看，数据集各观测点会落在一条曲线上。

（示例）函数关系

某种商品的销售额 Y与销售量 X之间的关系可表示为（ P为单价）：

$$Y_i = P_i \cdot X_i$$

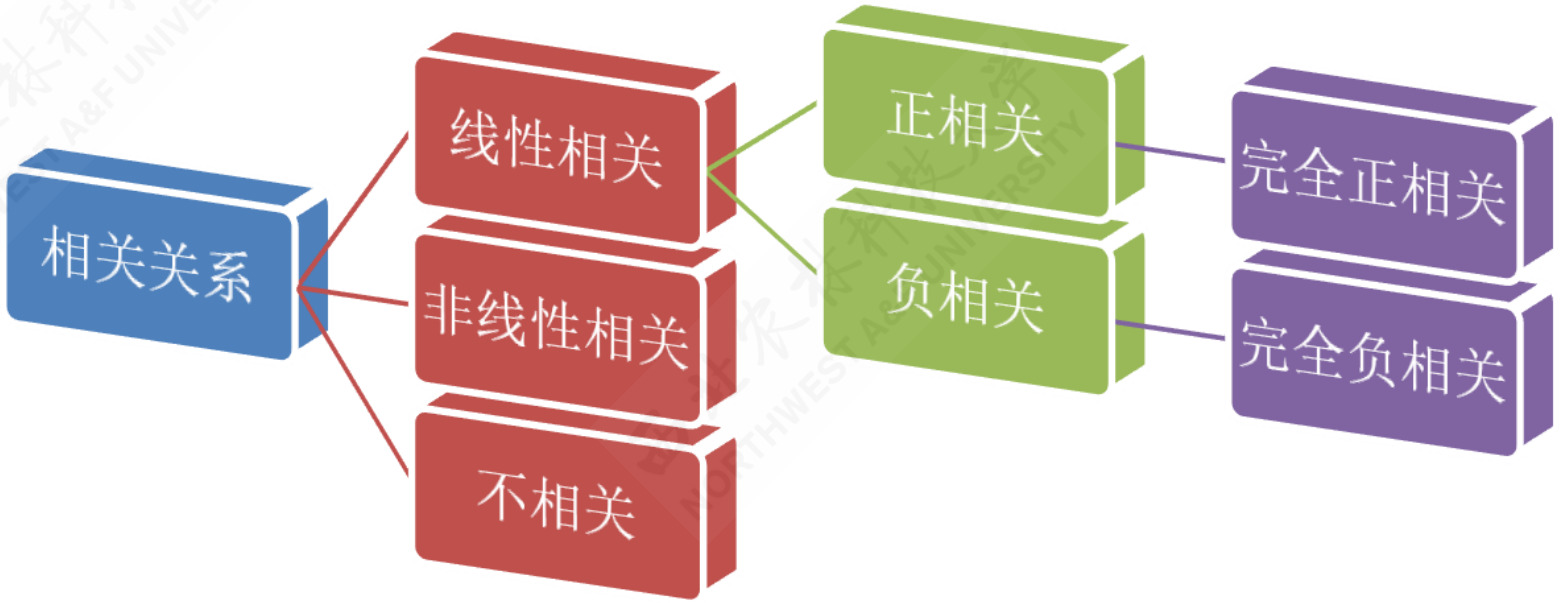
圆的面积 S与半径 R之间的关系可表示为：

$$S = \pi R^2$$

企业的原材料消耗额 Y与产量 $X1$ 、单位产量消耗 $X2$ 、原材料价格 $X3$之间的关系可表示为：

$$Y = X_1 \cdot X_2 \cdot X_3$$

变量间的关系：相关关系(correlation)



相关关系的类型

（示例）相关关系

- 父亲身高 Y与子女身高 X之间的关系
- 收入水平 Y与受教育程度 X之间的关系
- 粮食单位面积产量 Y与施肥量 $X1$ 、降雨量 $X2$、温度 $X3$之间的关系
- 商品的消费量 Y与居民收入 X之间的关系
- 商品销售额 Y与广告费支出 X之间的关系

相关关系的描述与测度：问题与假定

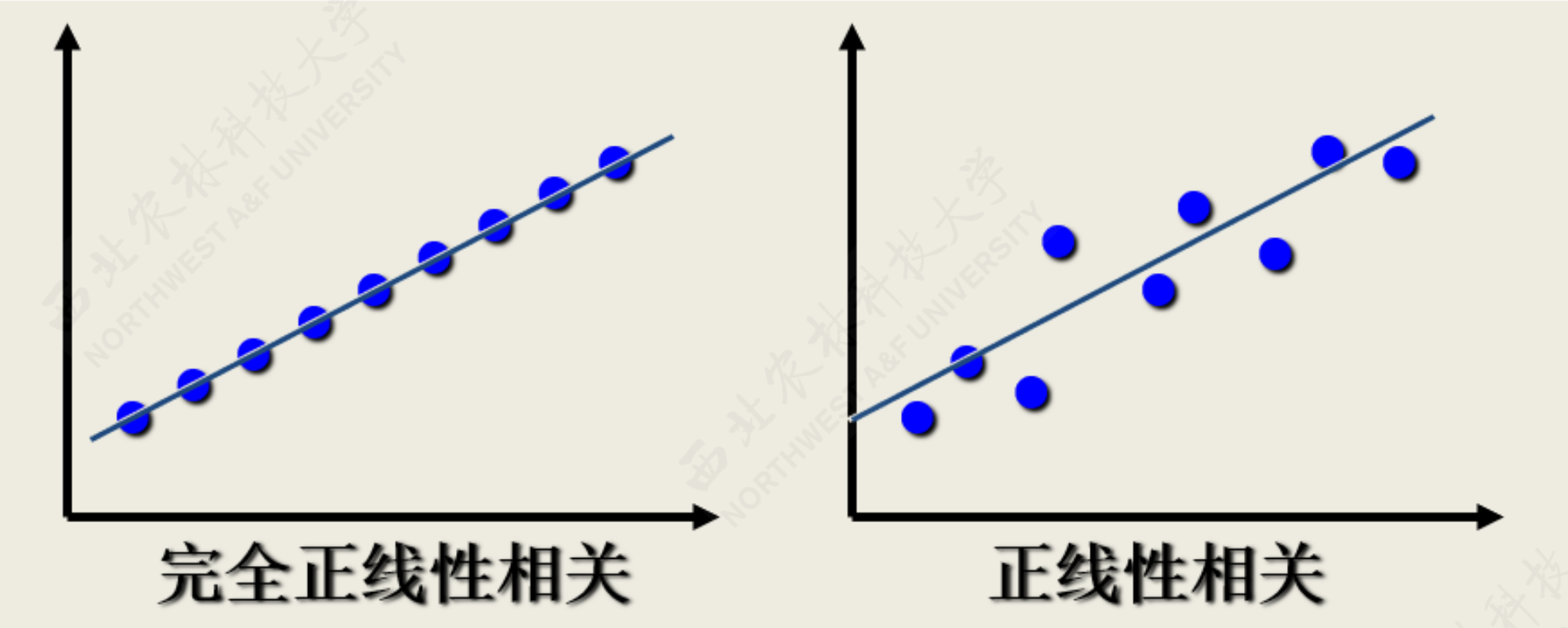
相关分析要解决的问题：

- 变量之间是否存在关系？
- 如果存在关系，它们之间是什么样的关系？
- 变量之间的关系强度如何？
- 样本所反映的变量之间的关系能否代表总体变量之间的关系？

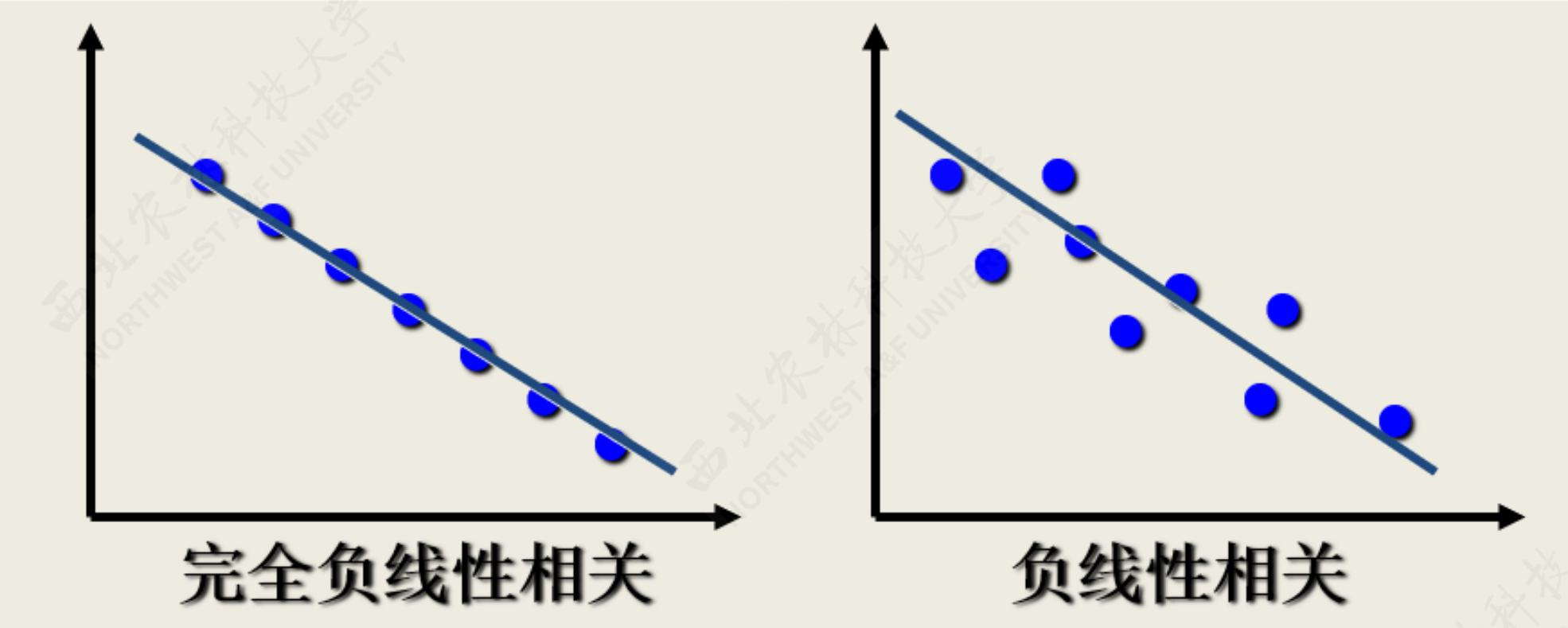
相关分析中的总体假定：

- 两个变量之间是线性关系
- 两个变量都是随机变量

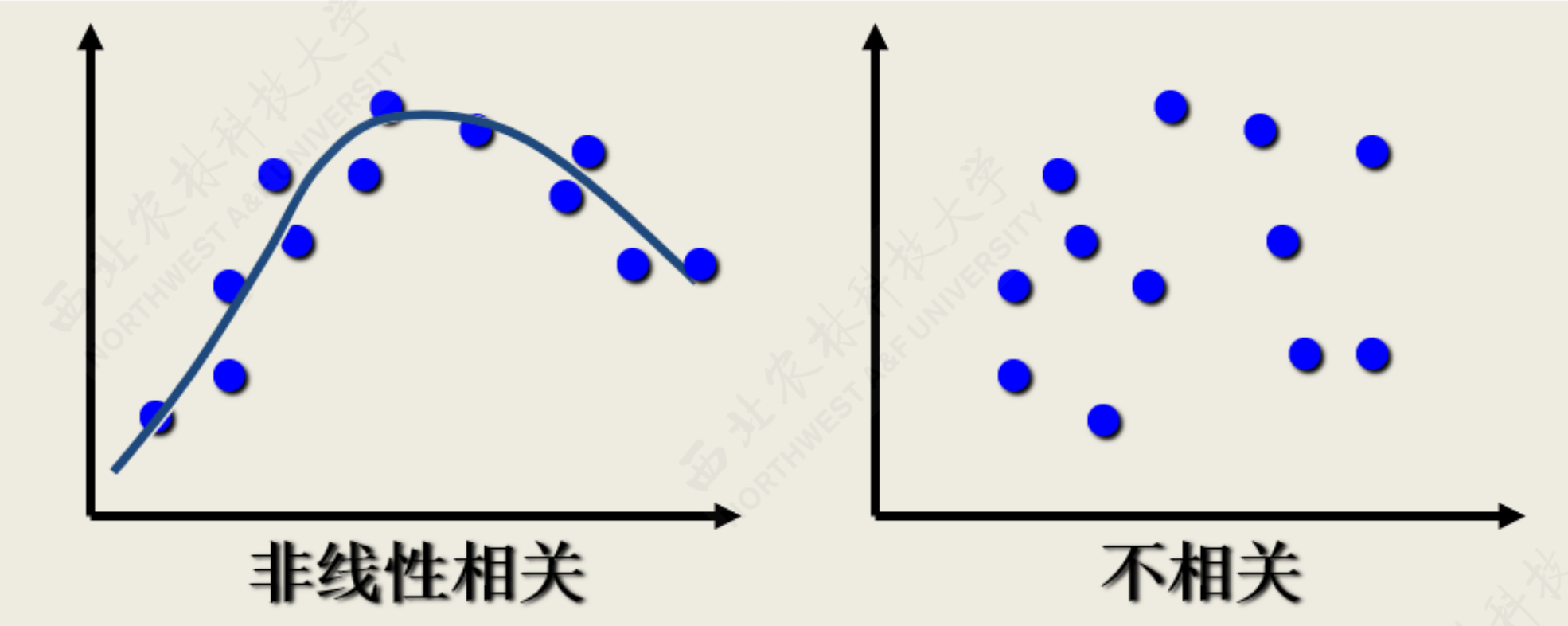
相关关系的描述与测度：散点图



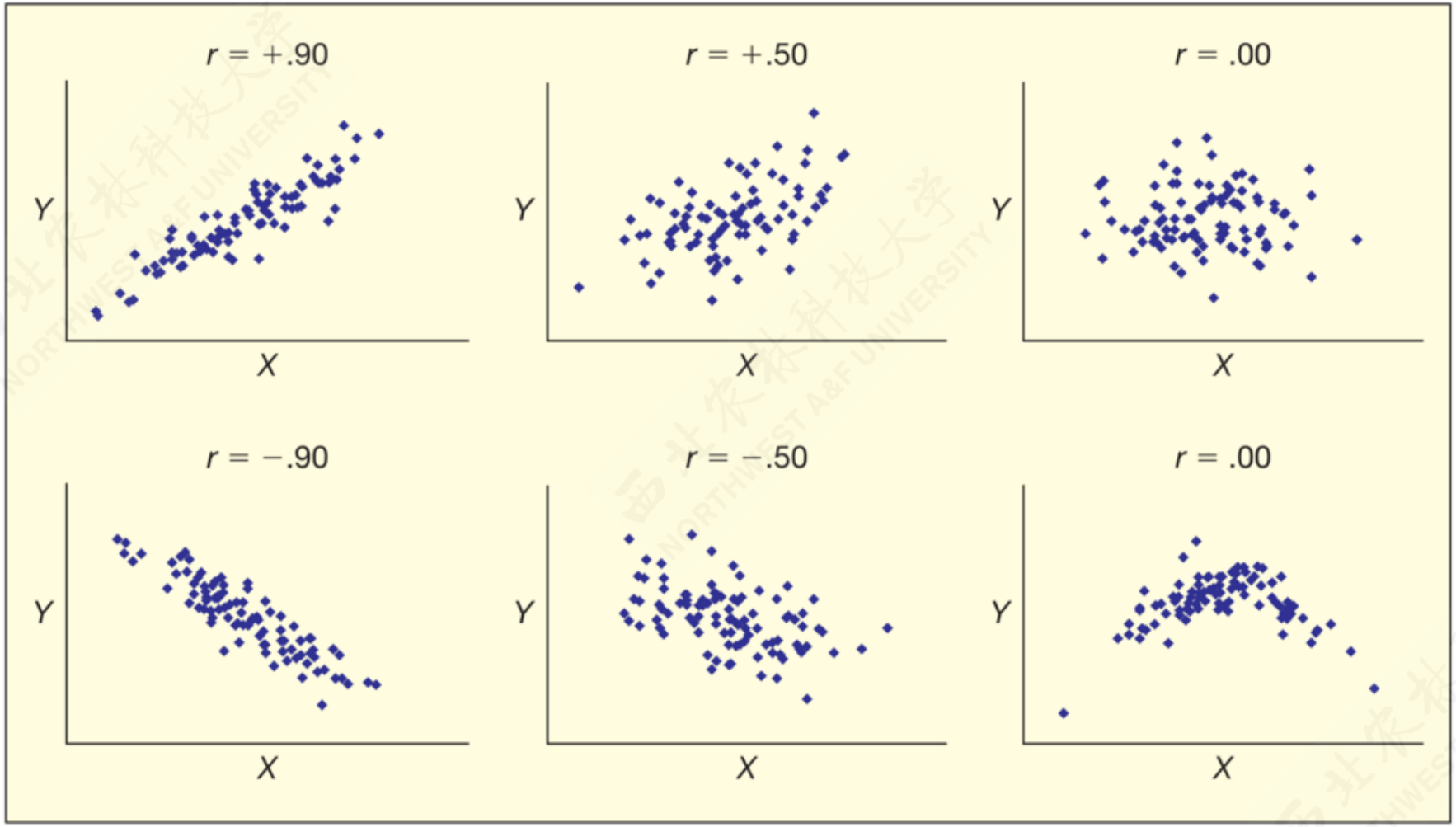
相关关系的描述与测度：散点图



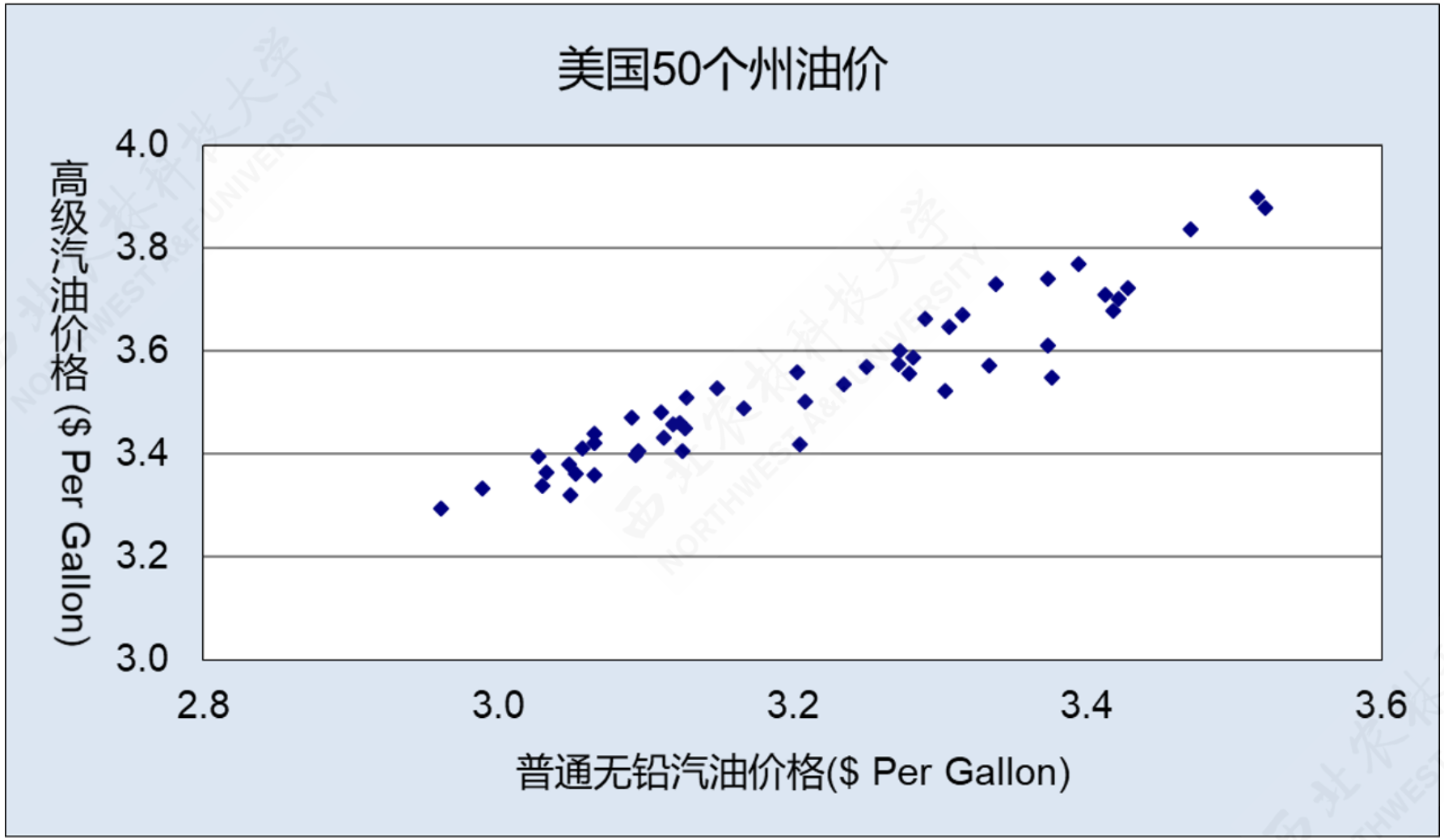
相关关系的描述与测度：散点图



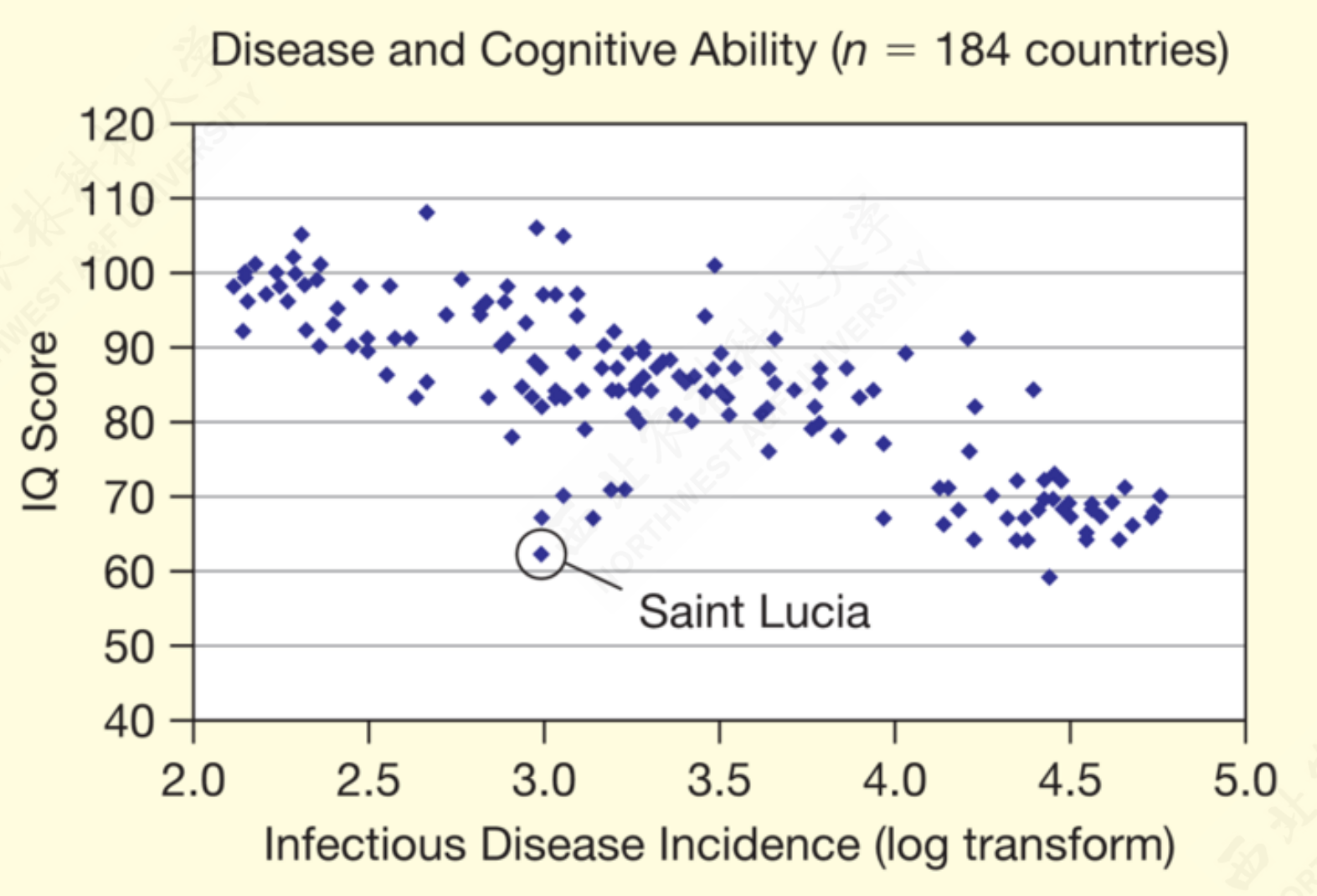
相关关系的描述与测度：散点图



（示例）两类油价的散点图


美国50个州油价
高级汽油价格($ Per Gallon)
4.0
3.8
3.6
3.4
3.2
3.0
2.8
3.0
3.2
3.4
3.6
普通无铅汽油价格($ Per Gallon)

（示例）传染病与认知水平的散点图



相关关系的描述与测度：相关系数

相关系数(correlation coefficient)：是度量变量之间关系强度的一个统计量。

- 它是对两个变量之间线性相关强度的一种度量。
- 一般称为简单相关系数，也称为线性相关系数(linear correlation coefficient)。
- 或称为Pearson相关系数(Pearson’s correlation coefficient) 。

相关系数记号表达：

- 若相关系数是根据总体全部数据计算的，称为总体相关系数，记为 ρ。
- 若是根据样本数据计算的，则称为样本相关系数，简称为相关系数，记为 r。

相关关系的描述与测度：计算公式

简单相关系数的大FF计算公式：

$$r=\frac{n\sum X_iY_i-\sum X_i\sum Y_i}{\sqrt{n\sum X_i^2-(\sum X_i)^2}\cdot\sqrt{n\sum Y_i^2-(\sum Y_i)^2}} \quad \text{(eq01)}$$

简单相关系数的小ff计算公式：

$$r=\frac{\sum\left((X_i-\overline{X})(Y_i-\overline{Y})\right)}{\sqrt{\sum(X_i-\overline{X})^2\sum(Y_i-\overline{Y})^2}}=\frac{SS_{XY}}{\sqrt{SS_{XX}}\sqrt{SS_{YY}}}=\frac{\sum x_iy_i}{\sqrt{\sum x_i^2\sum y_i^2}} \quad \text{(eq02)}$$

$$SS_{XX}=\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2;\quad SS_{YY}=\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2;\quad SS_{XY}=\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)$$

相关关系的描述与测度：特征

简单相关系数的特征：

性质1： r的取值范围是 $[-1,1]$， $|r|$越趋于1表示相关关系越强； $|r|$越趋于0表示相关关系越弱。

- 如果 $|r|=1$，为完全相关。其中 $r=1$，为完全正相关； $r=-1$，为完全负正相关
- 如果 $r=0$，不存在线性相关关系
- 如果 $-1<r<0$，为负相关；如果 $0<r<1$，为正相关。

性质2：r具有对称性。即 X与 Y之间的相关系数和 Y与 X之间的相关系数相等，即 $r_{XY}=r_{YX}$。

相关关系的描述与测度：特征

简单相关系数的特征：

性质3： r数值大小与 X和 Y原点及尺度无关，即改变 X和 Y的数据原点及计量尺度，并不改变 r数值大小。

性质4：仅仅是 X与 Y之间线性关系的一个度量，它不能用于描述非线性关系。这意为着， $r=0$只表示两个变量之间不存在线性相关关系，并不说明变量之间没有任何关系

性质5： r虽然是两个变量之间线性关系的一个度量，却不一定意味着 X与 Y一定有因果关系。

相关关系的描述与测度：解释

下面给出实证研究时，对相关系数的经验解释：

- 当 $|r| < 0.8$时，可视为两个变量之间高度相关。
- 当 $0.5 < |r| < 0.8$时，可视为中度相关。
- 当 $0.3 < |r| < 0.5$时，视为低度相关。
- 当 $|r| < 0.3$时，说明两个变量之间的相关程度极弱，可视为不相关。

而且上述解释必须建立在对相关系数的显著性进行检验的基础之上。

相关关系的描述与测度：简单相关系数

简单相关系数（simple correlation coefficient）：

- Y_i和 X_{2i}之间的相关系数：

$$r_{12}=\frac{\sum y_i x_{2i}}{\sqrt{\sum y_i^2}\sqrt{\sum x_{2i}^2}}$$

- Y_i和 X_{3i}之间的相关系数：

$$r_{13}=\frac{\sum y_i x_{3i}}{\sqrt{\sum y_i^2}\sqrt{\sum x_{3i}^2}}$$

- X_{2i}和 X_{3i}之间的相关系数：

$$r_{23}=\frac{\sum x_{2i} x_{3i}}{\sqrt{\sum x_{2i}^2}\sqrt{\sum x_{3i}^2}}$$

相关关系的描述与测度：偏相关系数

偏相关系数（partial correlation coefficient）： 一个不依赖于 X_{2i}的，对 X_{3i}和 Y_i的影响的一种相关系数。

- 保持 X_{3i}不变， Y_i和 X_{2i}之间的相关系数：

$$r_{12\cdot 3}=\frac{r_{12}-r_{13}r_{23}}{\sqrt{\left(1-r_{13}^2\right)\left(1-r_{23}^2\right)}}$$

- 保持 X_{2i}不变， Y_i和 X_{3i}之间的相关系数：

$$r_{13.2}=\frac{r_{13}-r_{12}r_{23}}{\sqrt{\left(1-r_{12}^2\right)\left(1-r_{23}^2\right)}}$$

- 保持 Y_i不变， X_{2i}和 X_{3i}之间的相关系数：

$$r_{23.1}=\frac{r_{23}-r_{12}r_{13}}{\sqrt{\left(1-r_{12}^2\right)\left(1-r_{13}^2\right)}}$$

相关系数的显著性检验

相关系数的显著性检验，是指检验两个变量之间是否存在线性相关关系。

相关系数的显著性检验方法包括：

- 等价于对回归斜率系数 β_1的检验（仅针对一元回归）
- 采用R. A. Fisher提出的t检验

相关系数的显著性检验

相关系数的显著性检验步骤：

1）提出假设： $H_0:\rho=0; H_1:\rho\neq 0$

2）计算样本统计量

$$T^*=|r|\sqrt{\frac{n-2}{1-r^2}} \sim t(n-2)$$

3）给定显著性水平 α，确定t理论分布值 $t_{1-\alpha/2}(n-2)$。

4）得到假设检验结论：

- 若 $T^*>t_{1-\alpha/2}(n-2)$，则拒绝 H_0，认为显著存在相关关系；
- 若 $T^*<t_{1-\alpha/2}(n-2)$，则无法拒绝 H_0，认为相关关系不显著。

（案例）银行贷款：案例数据

案例说明：某银行共有25家分行，分行及所在地区的相关变量数据如下表所示。

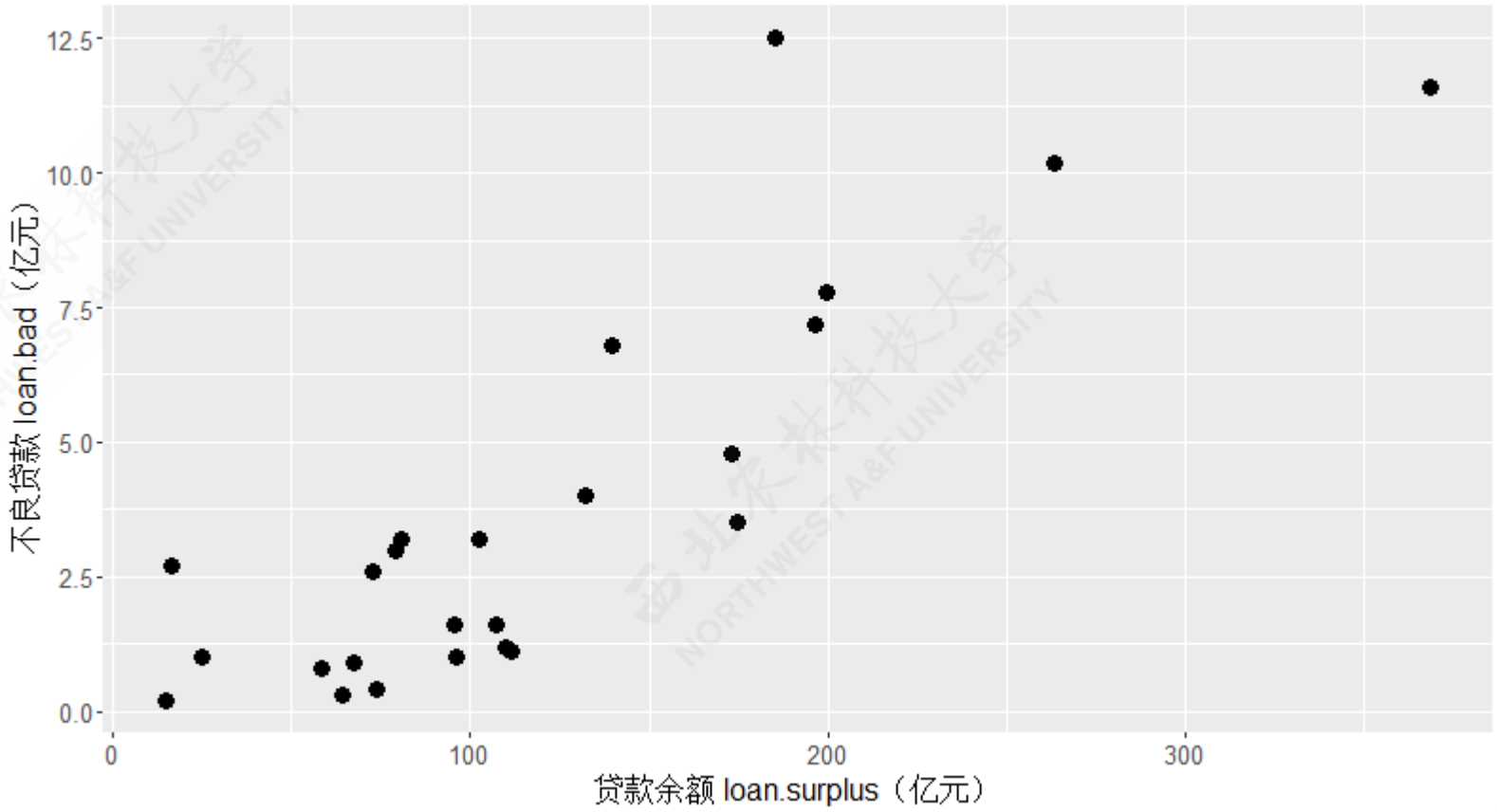
ID.bank	loan.bad	loan.surplus	loan.receivable	loan.numbers	invest
1	0.9	67.3	6.8	5	
2	1.1	111.3	19.8	16	
3	4.8	173	7.7	17	
4	3.2	80.8	7.2	10	
5	7.8	199.7	16.5	19	
6	2.7	16.2	2.2	1	
7	1.6	107.4	10.7	17	

Showing 1 to 7 of 25 entries

Previous 1 2 3 4 Next

说明：上述变量的含义分别是ID.bank（分行编号）、loan.bad（不良贷款）、loan.surplus（各项贷款余额 ）、loan.receivable（本年累计应收贷款）、loan.numbers（贷款项目个数）、investment.fixed（本年固定资产投资额）。

（案例）银行贷款：不良贷款VS贷款余额的散点图



不良贷款VS贷款余额散点图

（案例）银行贷款：不良贷款VS贷款余额的相关系数（大数）

大数计算表

ID. bank	Y	X	XY	X_sqr	Y_sqr
1	0.9	67.3	60.57	4,529.29	0.81
2	1.1	111.3	122.43	12,387.69	1.21
3	4.8	173	830.40	29,929.00	23.04
4	3.2	80.8	258.56	6,528.64	10.24
5	7.8	199.7	1,557.66	39,880.09	60.84
6	2.7	16.2	43.74	262.44	7.29
7	1.6	107.4	171.84	11,534.76	2.56

Showing 1 to 7 of 26 entries

Previous 1 2 3 4 Next

（案例）银行贷款：不良贷款VS贷款余额的相关系数（大FF）

相关系数 r的大FF计算公式（`eq01`）：

$$
\begin{aligned}
r &= \frac{n\sum X_iY_i - \sum X_i \sum Y_i}{\sqrt{n\sum X_i^2 - \left(\sum X_i\right)^2} \cdot \sqrt{n\sum Y_i^2 - \left(\sum Y_i\right)^2}} \\
&= \frac{25 \times 17080.14 - 3006.7 \times 93.2}{\sqrt{25 \times 516543.37 - (3006.7)^2} \cdot \sqrt{25 \times 660.1 - (93.2)^2}} \\
&= 0.8436
\end{aligned}
$$

（案例）银行贷款：不良贷款VS贷款余额的相关系数（小的）

小的计算表

ID. bank	Y	X	x	y	x_sqr	y_sqr	xy
1	0.9	67.3	-52.97	-2.83	2,805.61	8.00	149.79
2	1.1	111.3	-8.97	-2.63	80.43	6.91	23.57
3	4.8	173	52.73	1.07	2,780.66	1.15	56.53
4	3.2	80.8	-39.47	-0.53	1,557.72	0.28	20.84
5	7.8	199.7	79.43	4.07	6,309.44	16.58	323.45
6	2.7	16.2	-104.07	-1.03	10,830.15	1.06	106.98
7	1.6	107.4	-12.87	-2.13	165.59	4.53	27.38

Showing 1 to 7 of 26 entries

Previous 1 2 3 4 Next

相关系数 r的小FF计算公式（`eq02`）：

$$
\begin{aligned}
r &= \frac{\sum\left((X_i-\overline{X})(Y_i-\overline{Y})\right)}{\sqrt{\sum(X_i-\overline{X})^2(Y_i-\overline{Y})^2}} \\
&= \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}} \\
&= \frac{5871.16}{\sqrt{154933.57 \times 312.65}} \\
&= 0.8436
\end{aligned}
$$

（案例）银行贷款：相关系数矩阵表(Pearson)

```
corl_pearson<- round(cor(df_loan[,-1], method = "pearson"),4)
corl_pearson[upper.tri(corl_pearson)]<- NA
```

Pearson相关系数矩阵

	loan.bad	loan.surplus	loan.receivable	loan.numbers	investment.fixed
loan.bad	1.0000				
loan.surplus	0.8436	1.0000			
loan.receivable	0.7315	0.6788	1.0000		
loan.numbers	0.7003	0.8484	0.5858	1.0000	
investment.fixed	0.5185	0.7797	0.4724	0.7466	1.0000

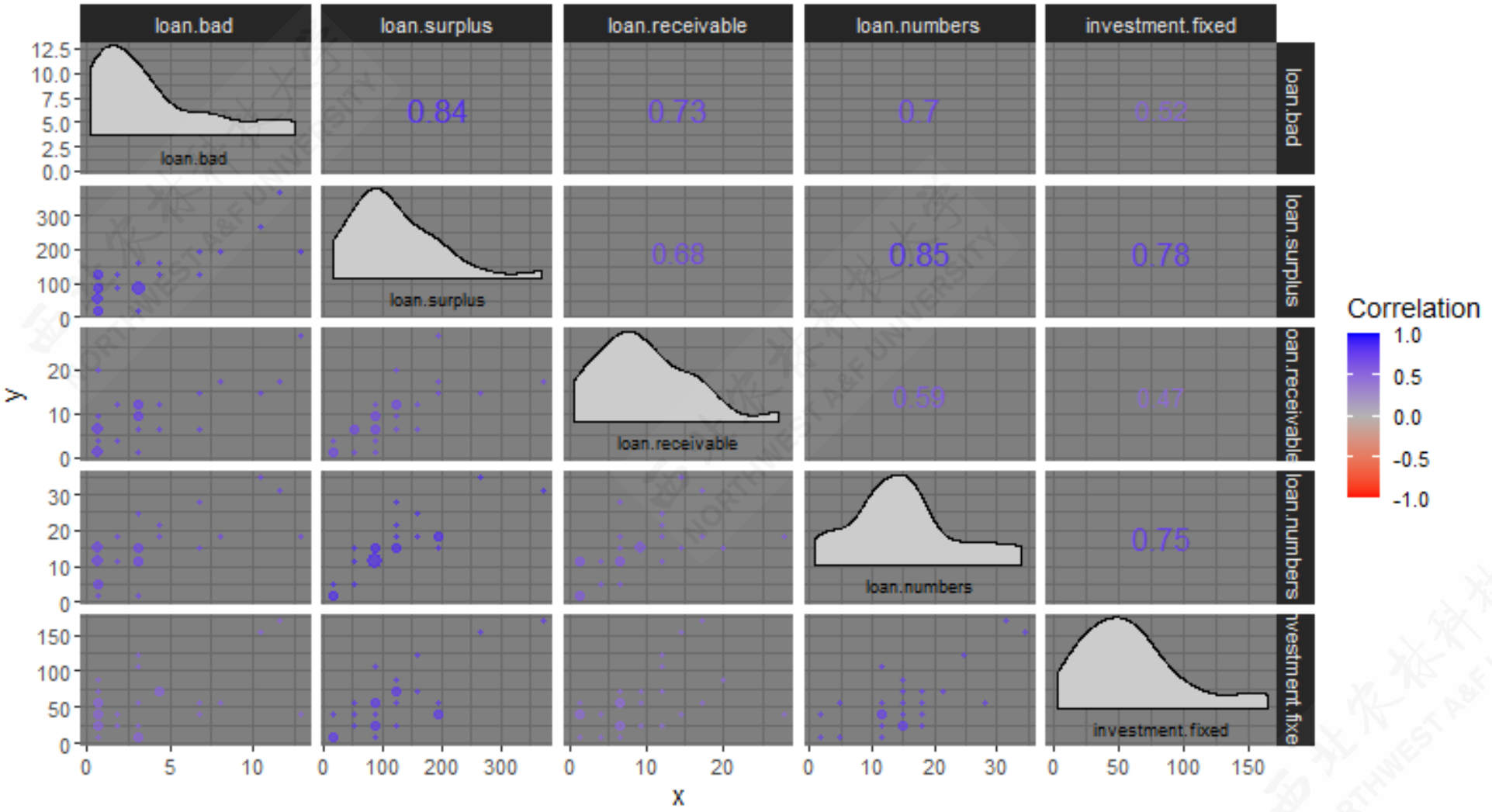
（案例）银行贷款：相关系数矩阵(Spearman)

```
corl_spearman<- round(cor(df_loan[,-1], method = "spearman"),4)
corl_spearman[upper.tri(corl_spearman)] <- NA
```

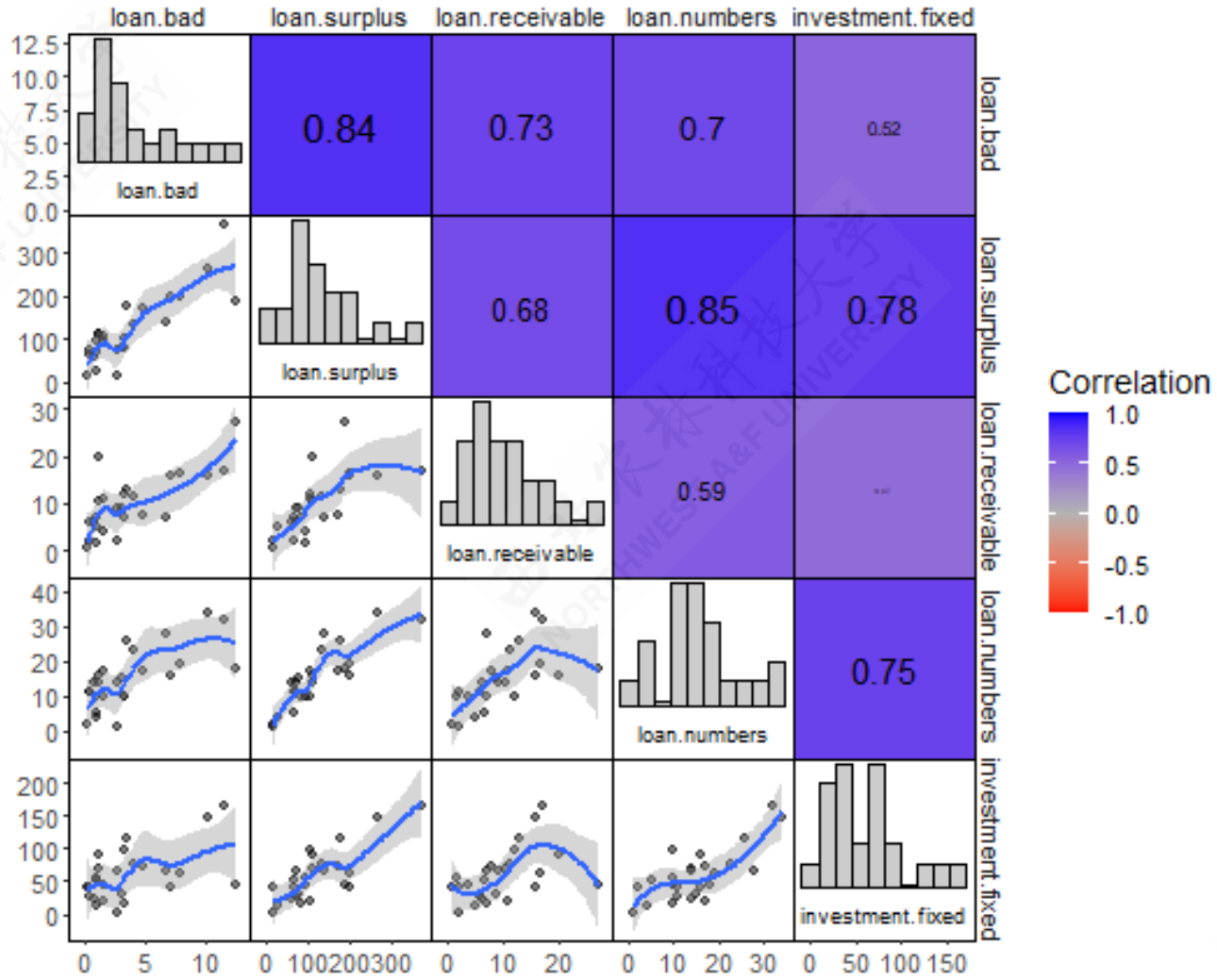
Spearman相关系数矩阵

	loan.bad	loan.surplus	loan.receivable	loan.numbers	investment.fixed
loan.bad	1.0000				
loan.surplus	0.8339	1.0000			
loan.receivable	0.7331	0.8148	1.0000		
loan.numbers	0.7172	0.8559	0.7393	1.0000	
investment.fixed	0.4407	0.6582	0.5469	0.5975	1.0000

（案例）银行贷款：相关系数矩阵图



（案例）银行贷款：相关系数矩阵图



（案例）银行贷款：偏相关系数

假定我们认为不良贷款（`loan.bad`）与贷款余额（`loan.surplus`）及贷款项目数（`loan.number`）存在相互关系。

前面我们已经计算出如下的简单相关系数：

$$r_{12} = r_{bad,sur} = 0.8436; \quad r_{13} = r_{bad,num} = 0.7003; \quad r_{23} = r_{num,sur} = 0.8484$$

因此我们可以分别计算出偏相关系数

（案例）银行贷款：偏相关系数

- 保持 X_{3i}不变， Y_i和 X_{2i}之间的相关系数：

$$r_{12\cdot 3}=\frac{r_{12}-r_{13}r_{23}}{\sqrt{\left(1-r_{13}^2\right)\left(1-r_{23}^2\right)}}=\frac{0.84-0.7\times 0.85}{\sqrt{\left(1-0.7^2\right)\left(1-0.85^2\right)}}=0.6601$$

- 保持 X_{2i}不变， Y_i和 X_{3i}之间的相关系数：

$$r_{13.2}=\frac{r_{13}-r_{12}r_{23}}{\sqrt{\left(1-r_{12}^2\right)\left(1-r_{23}^2\right)}}=\frac{0.7-0.84\times 0.85}{\sqrt{\left(1-0.84^2\right)\left(1-0.85^2\right)}}=-0.0542$$

- 保持 Y_i不变， X_{2i}和 X_{3i}之间的相关系数：

$$r_{23.1}=\frac{r_{23}-r_{12}r_{13}}{\sqrt{\left(1-r_{12}^2\right)\left(1-r_{13}^2\right)}}=\frac{0.85-0.84\times 0.7}{\sqrt{\left(1-0.84^2\right)\left(1-0.7^2\right)}}=0.6722$$

（案例）银行贷款：相关系数显著性检验(手算)

对于前述`loan.surplus`与`loan.bad`进行相关系数显著性检验（Pearson）：

- 1）提出假设： $H_0:\rho=0; H_1:\rho\neq 0$
- 2）计算样本统计量：

$$T^*=|r|\sqrt{\frac{n-2}{1-r^2}}=0.84\times\sqrt{\frac{25-2}{1-0.84^2}}=7.53$$

- 3）给定显著性水平 $\alpha=0.05$，确定t理论分布值
  $t_{1-\alpha/2}(n-2)=t_{1-0.05/2}(25-2)=t_{0.975}(23)=2.07$。
- 4）得到假设检验结论：因为t样本统计量大于t理论查表值，也即

$$[T^*=7.53]>[t_{0.975}(23)=2.07]$$

因此拒绝原假设 H_0，认为变量`loan.surplus`（贷款余额）与`loan.bad`（不良贷款）显著存在相关关系。

（案例）银行贷款：相关系数显著性检验(R软件)

我们可以使用R软件函数`cor.test()`对上述两个变量进行相关系数显著性检验：

```
cor.test(df_rel1$loan.surplus, df_rel1$loan.bad,
         method = "pearson")
```

```
    Pearson's product-moment correlation

data:  df_rel1$loan.surplus and df_rel1$loan.bad
t = 8, df = 23, p-value = 0.0000001
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.67 0.93
sample estimates:
 cor
0.84
```

本节结束