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

2022-02-22

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

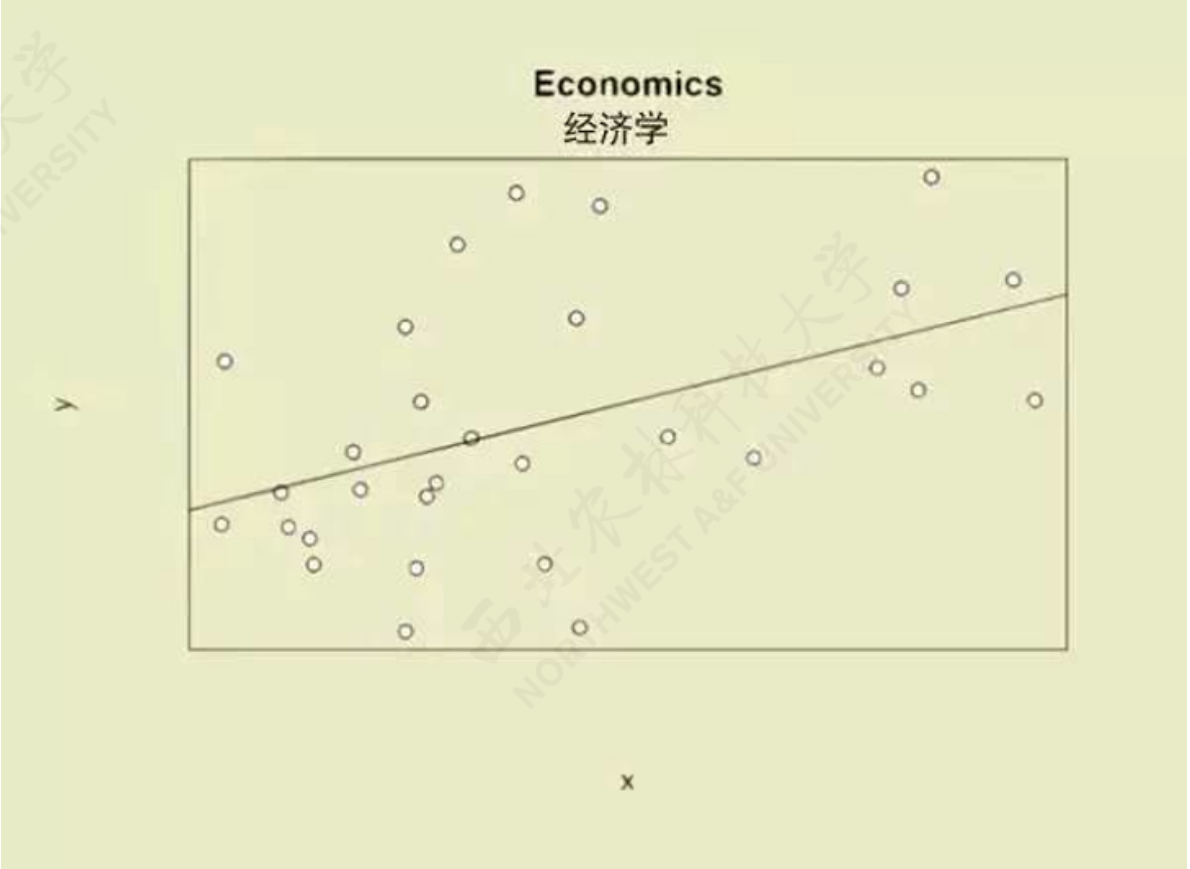
5.1 变量间关系的度量

变量间的关系

相关关系的描述与测度

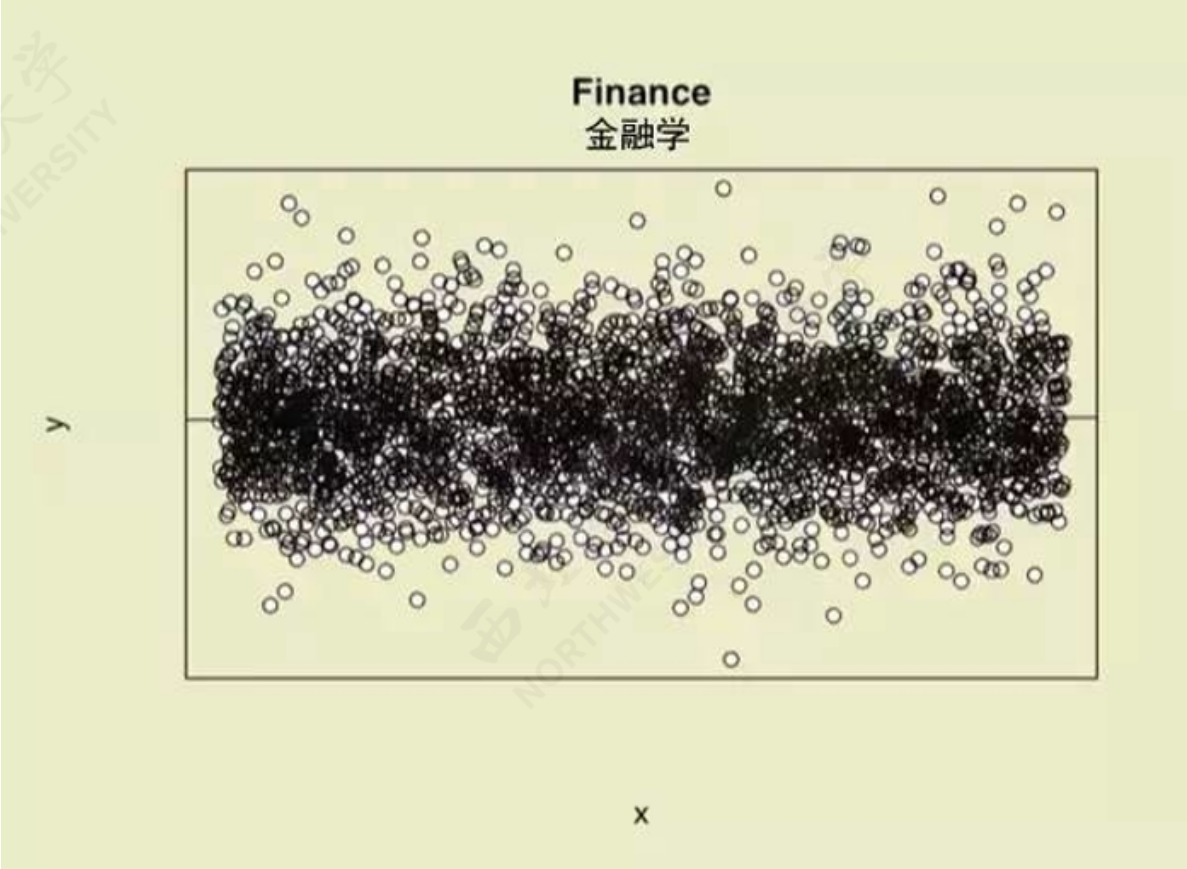
相关系数的显著性检验

（示例）变量间的关系：经济学专业解读



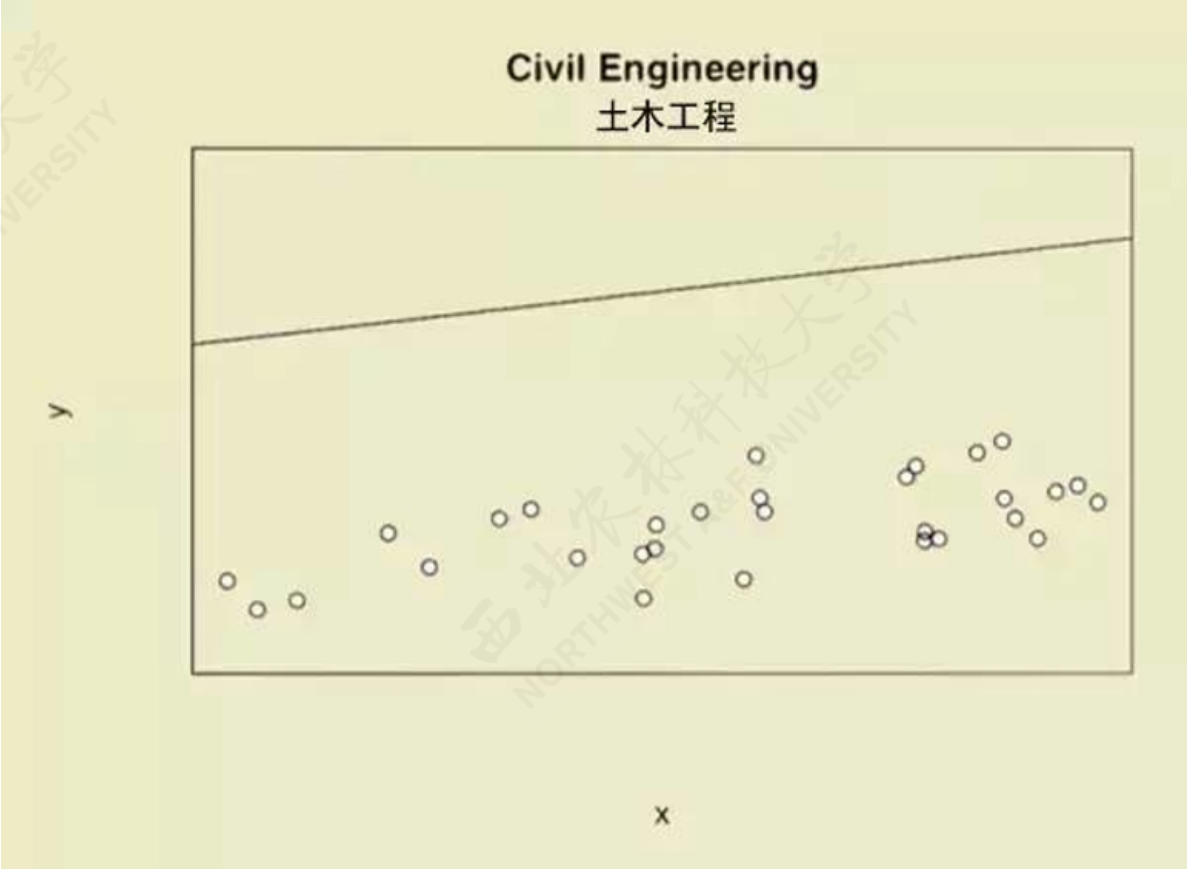
> “我们数据不少，做了很严格的回归，但异常值略多略多，符合理论的数值反而难找……”

（示例）变量间的关系：金融学专业解读



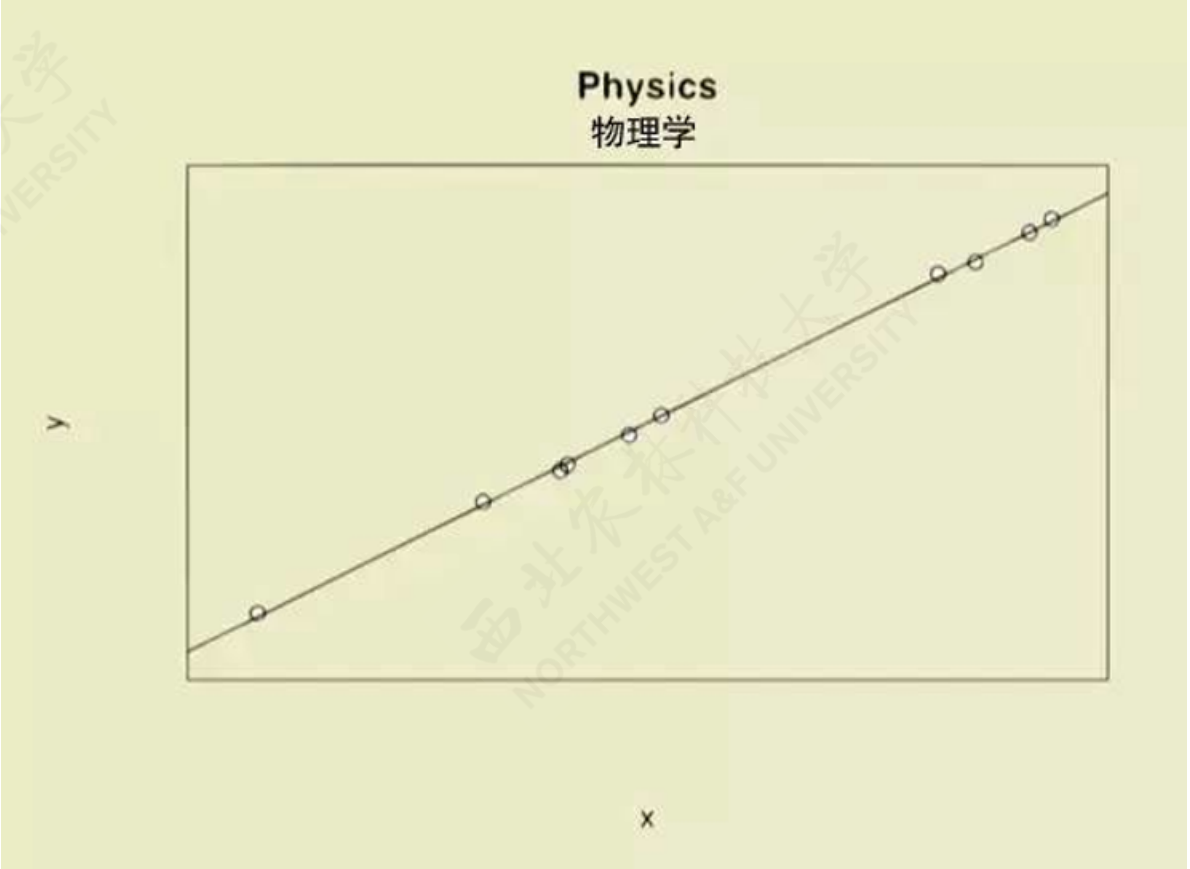
> “我们的数据多如牛毛，无孔不入。即使做完回归，也会发现异常值和符合理论的数值多得不忍直视。”

（示例）变量间的关系：土木工程专业解读



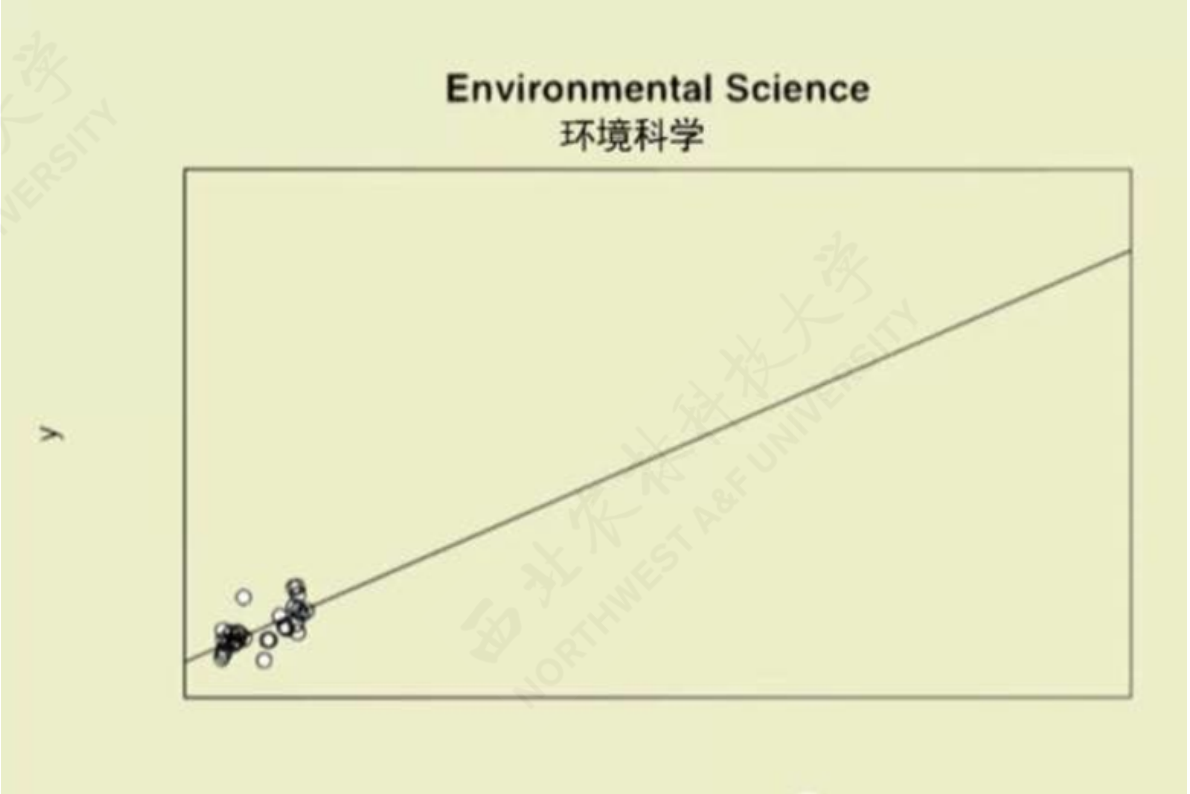
> “我们得要设计余量，所以理论设计得远高于实际承受……”

（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

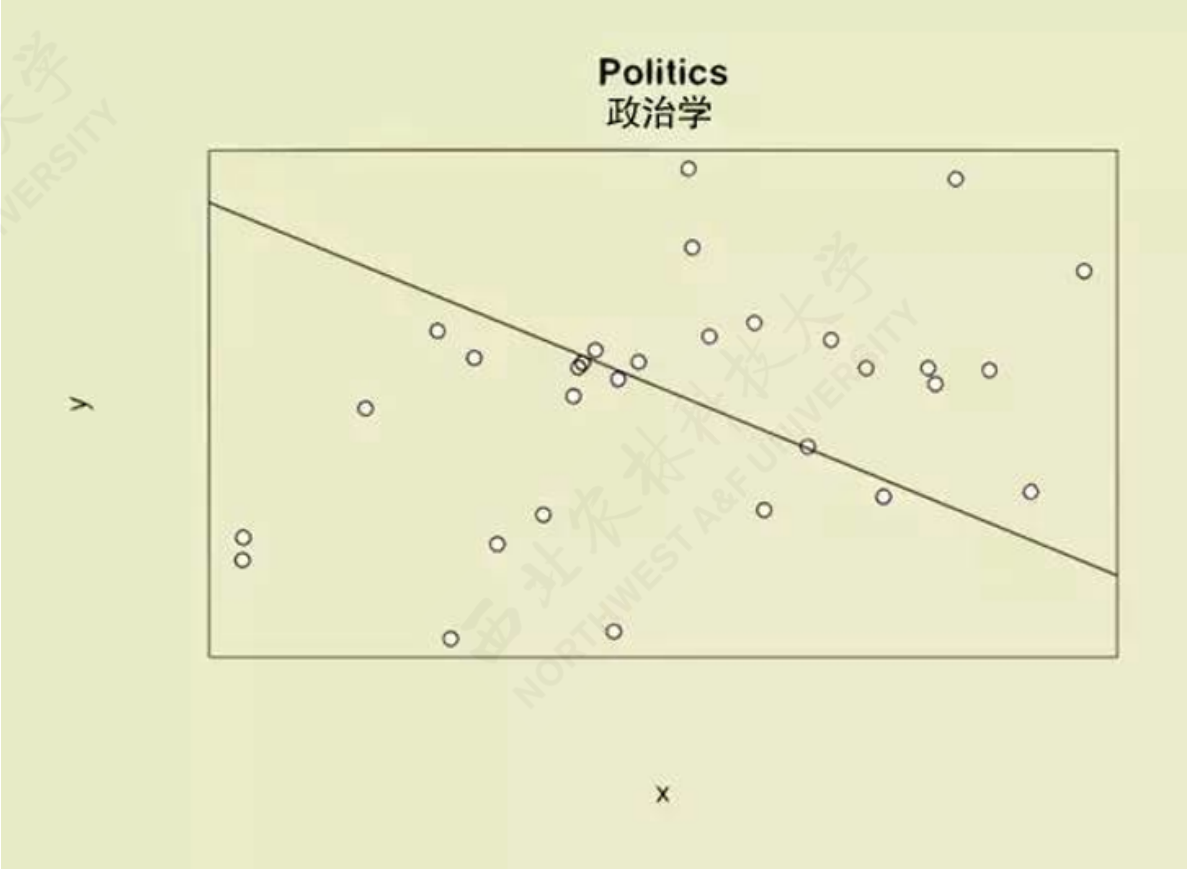


> “我们的理论和数据大致吻合，就是……应用范围有点蛋疼。”

（示例）变量间的关系：历史学专业解读

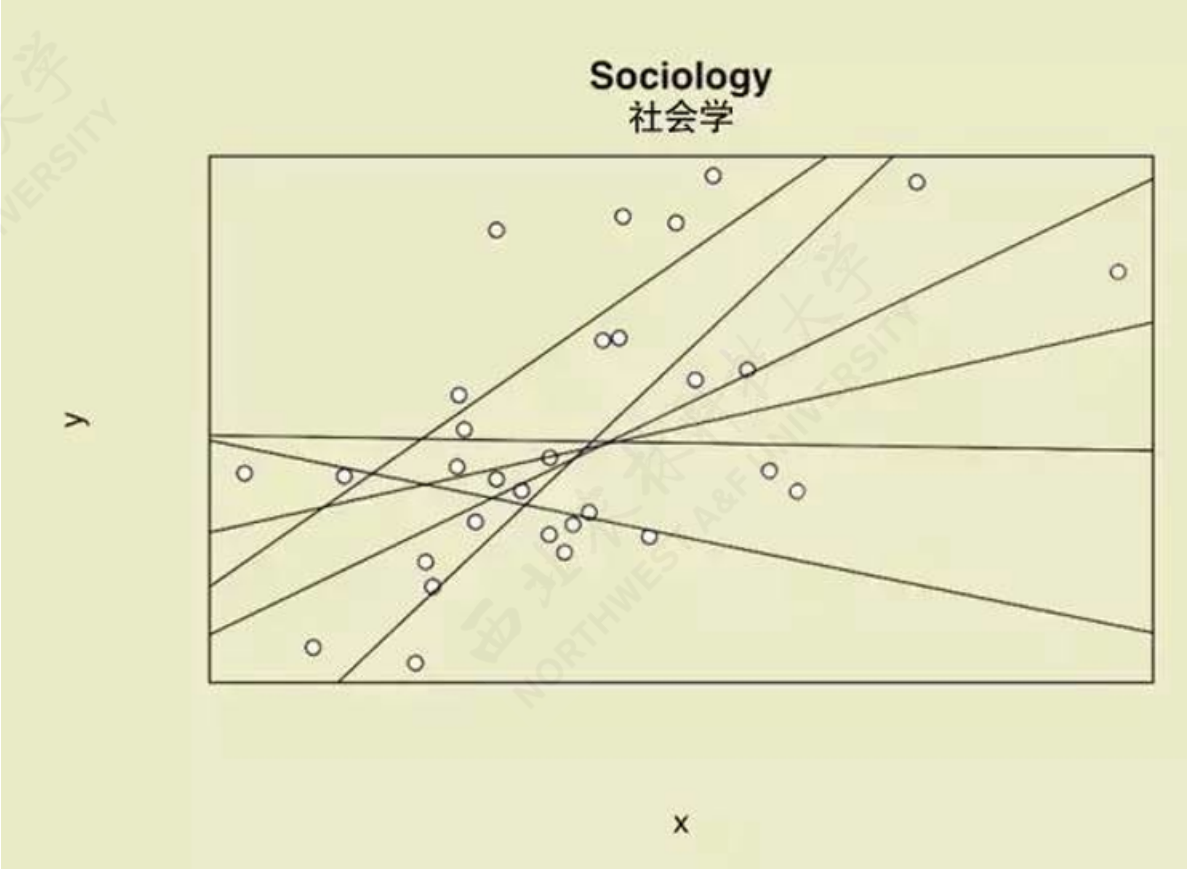
> “数据虽然很多，可我们能用理论把他们统统连起来！”

（示例）变量间的关系：政治学专业解读



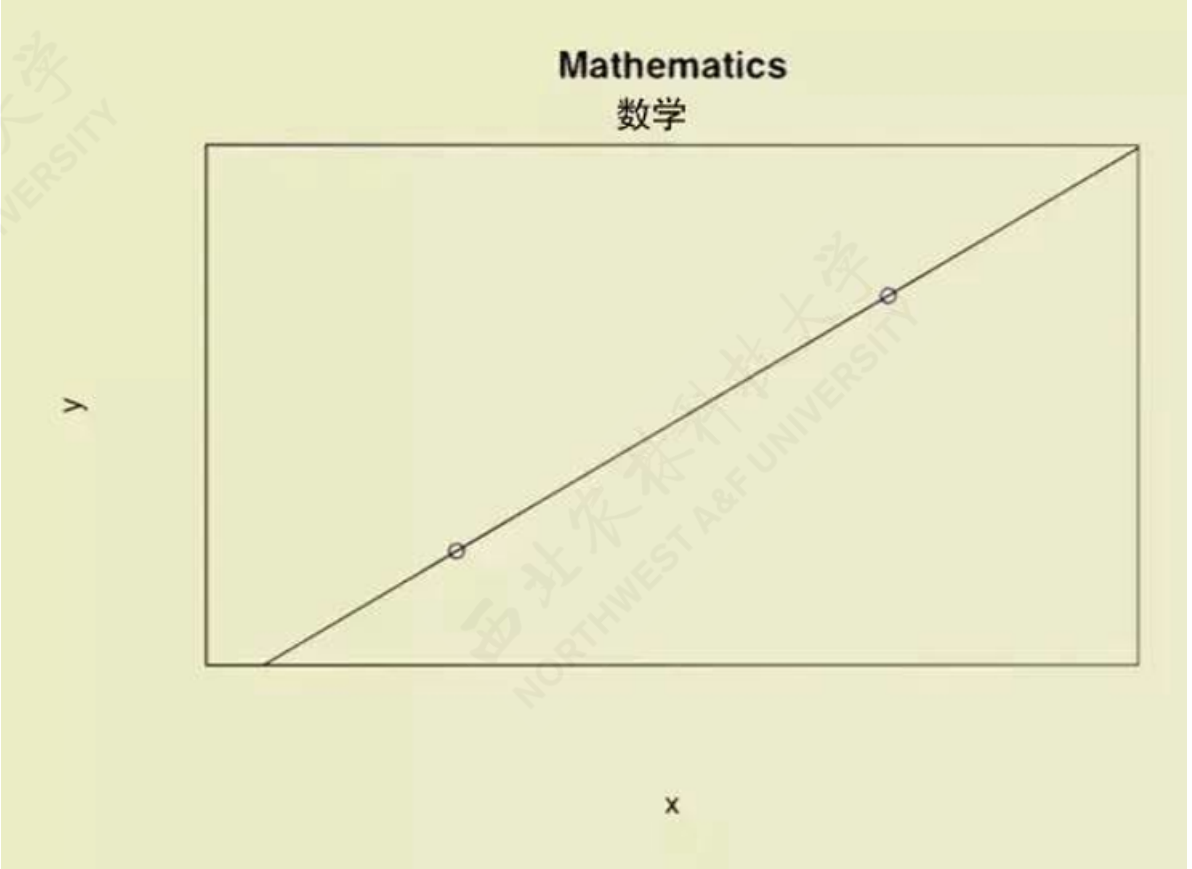
> “世界大势一日三变，尽管我们数据不少，可……我们的理论跟数据趋势是反着来的……”

（示例）变量间的关系：社会学专业解读



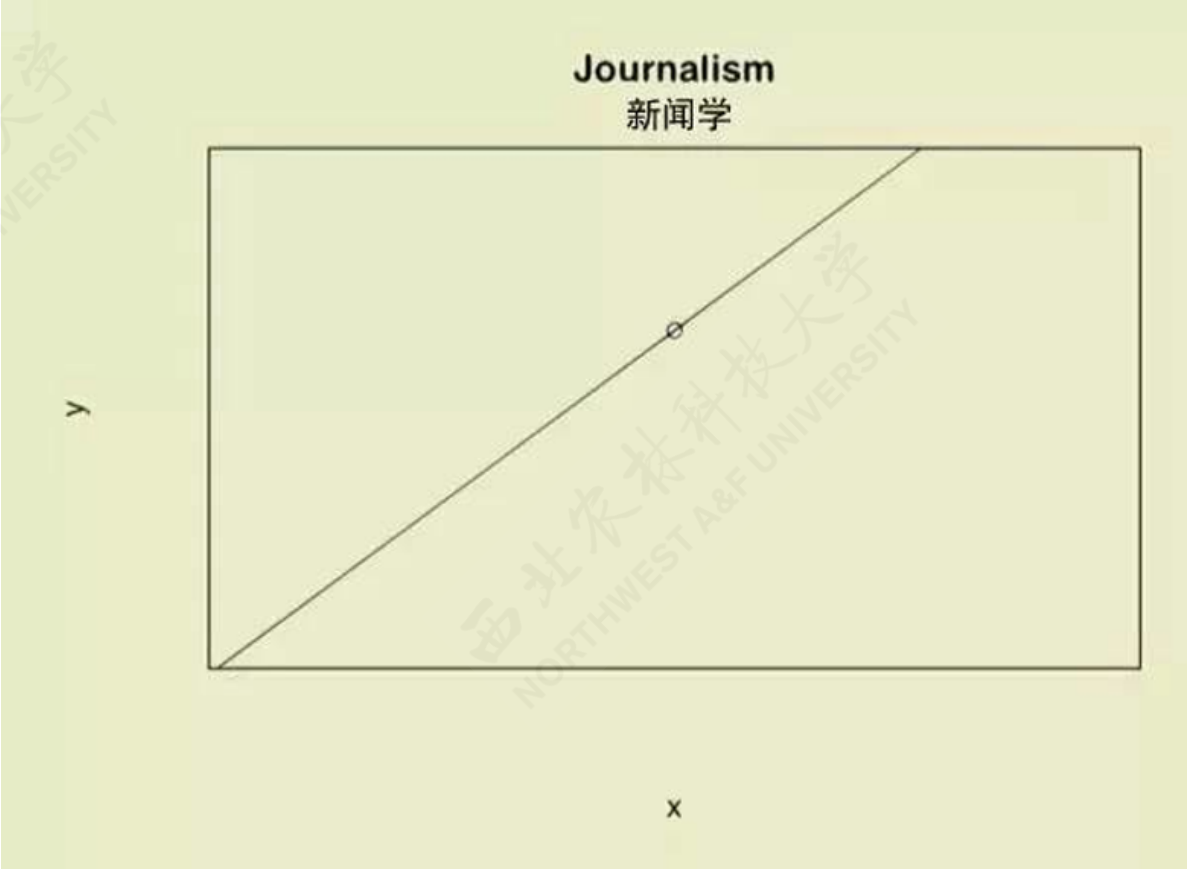
> “学海无涯苦作舟。那么多数据，那么多理论，慢慢学，恩……”

（示例）变量间的关系：数学专业解读



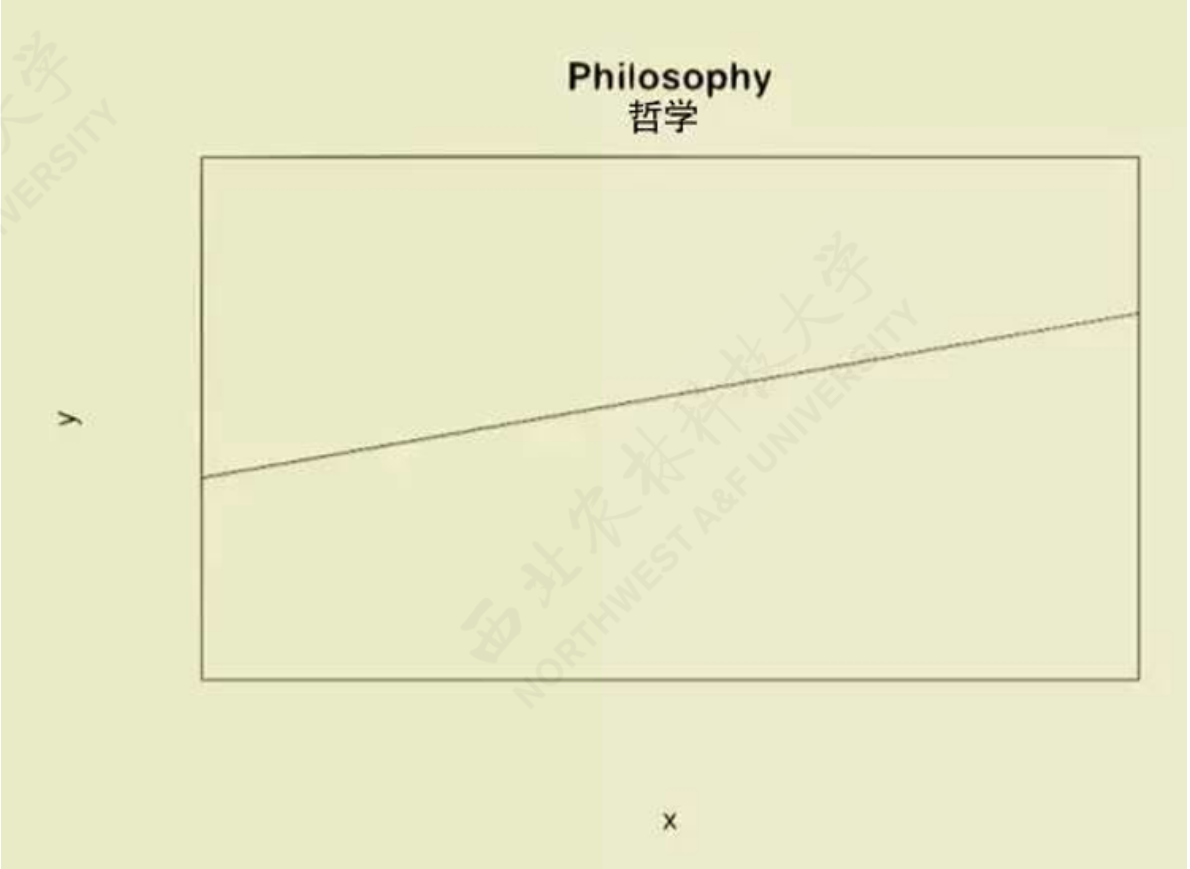
> “数据很少，但能建立理论～”

（示例）变量间的关系：新闻学专业解读



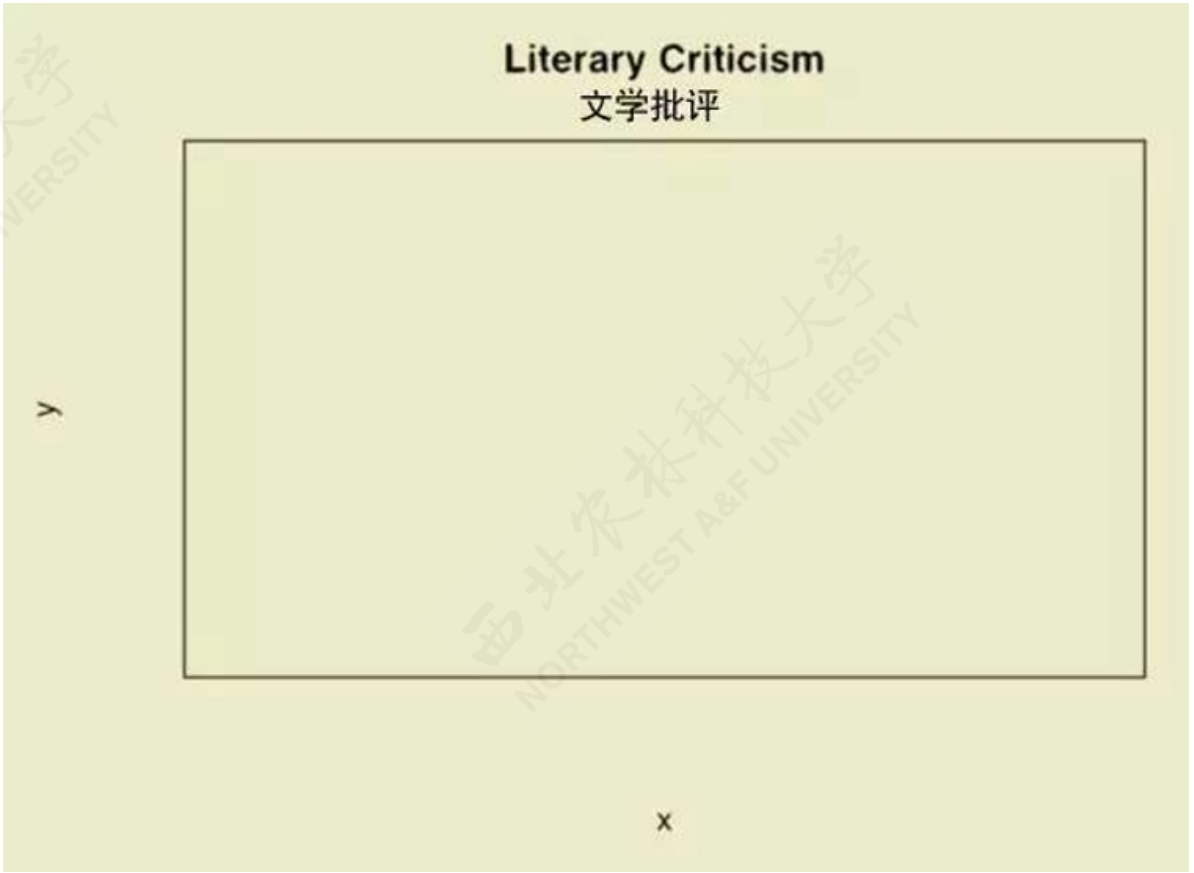
（示例）“只有一个数据，也能建立理论……”

（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)：

$$Y_i = P_i \cdot X_i$$

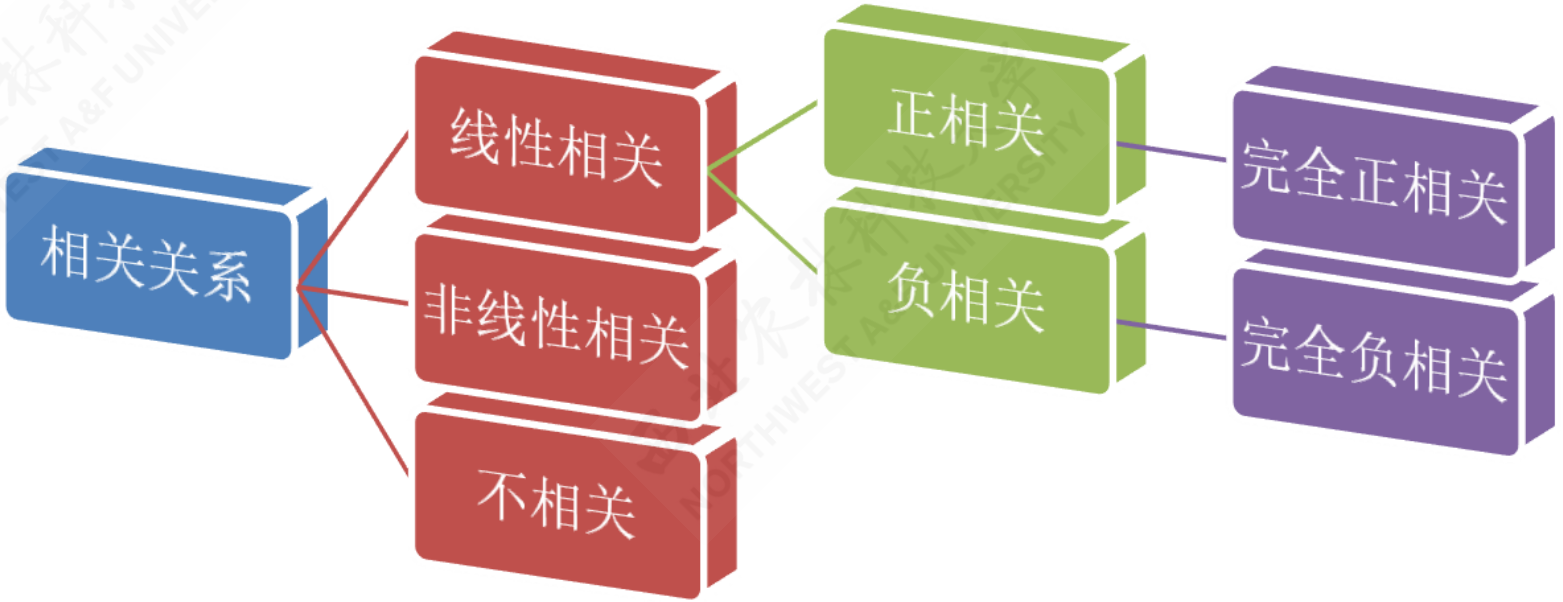
圆的面积S与半径R之间的关系可表示为：

$$S = \pi R^2$$

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

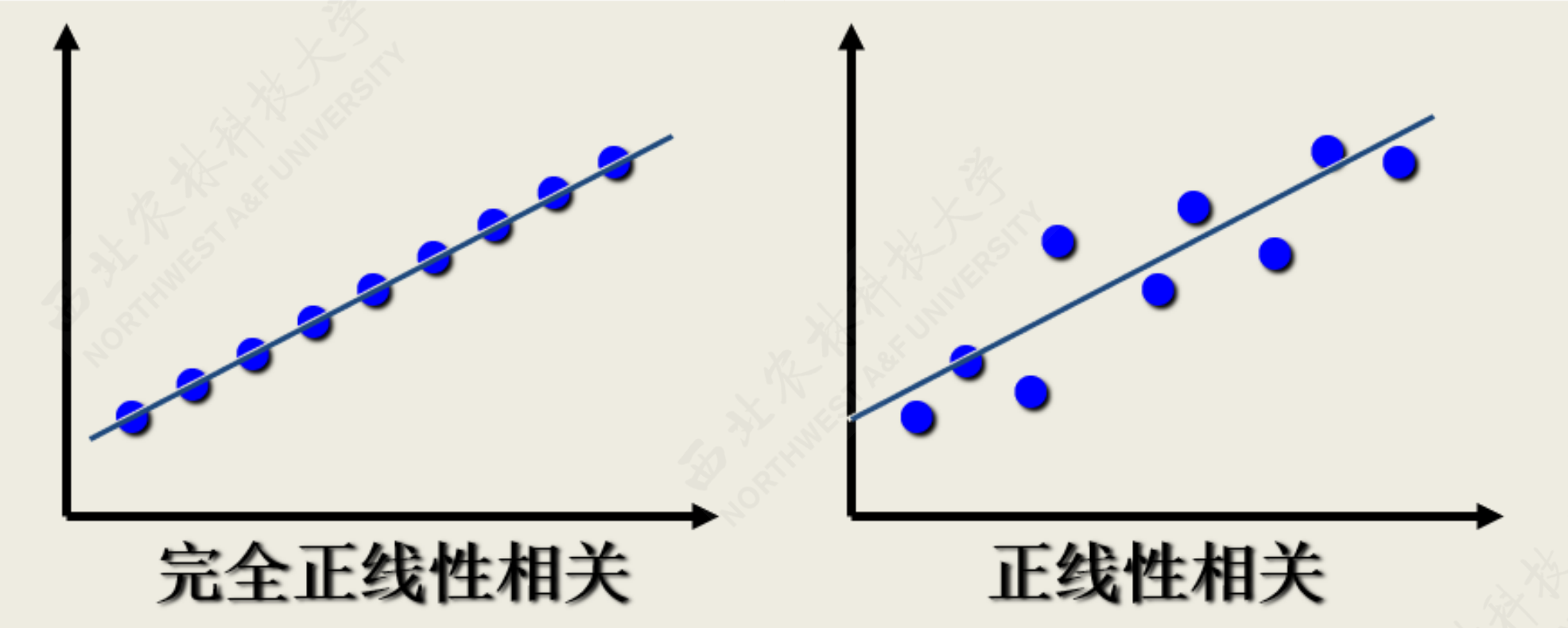
相关分析要解决的问题：

- 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？

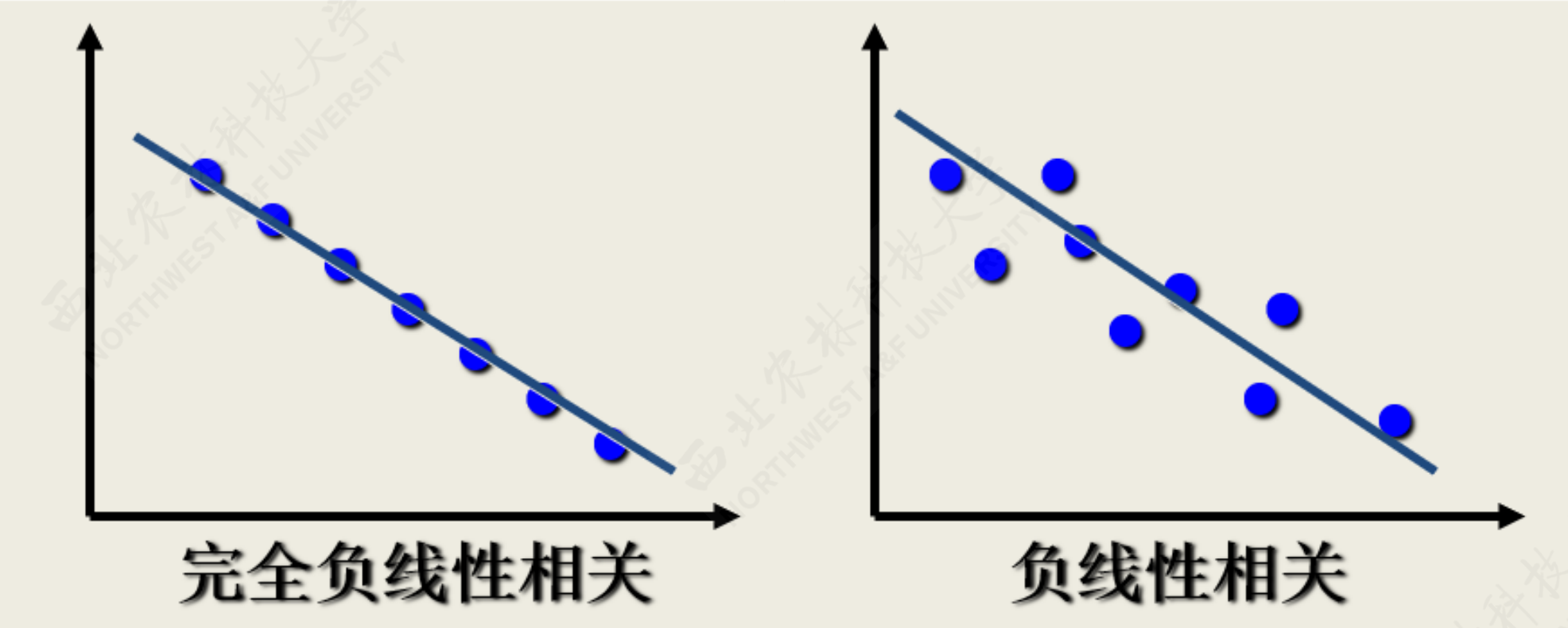
相关分析中的总体假定：

- 两个变量之间是线性关系
- 两个变量都是随机变量

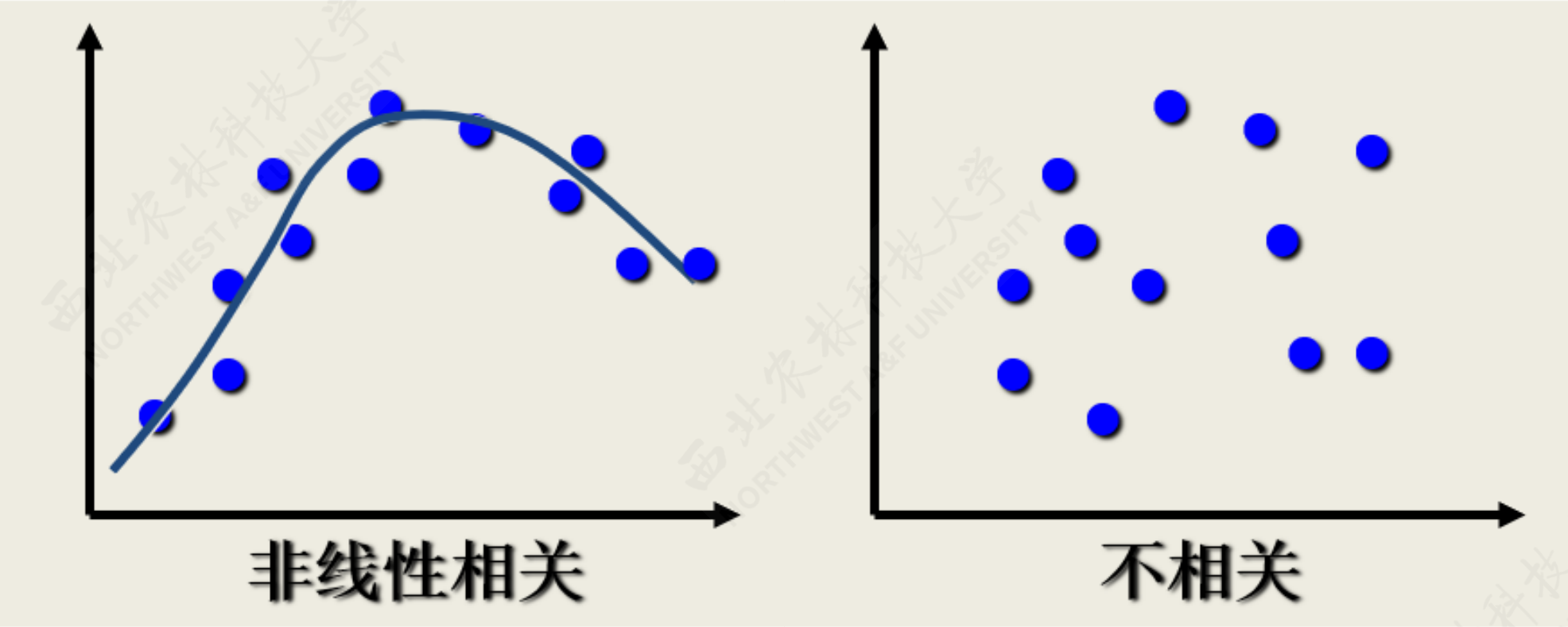
相关关系的描述与测度：散点图



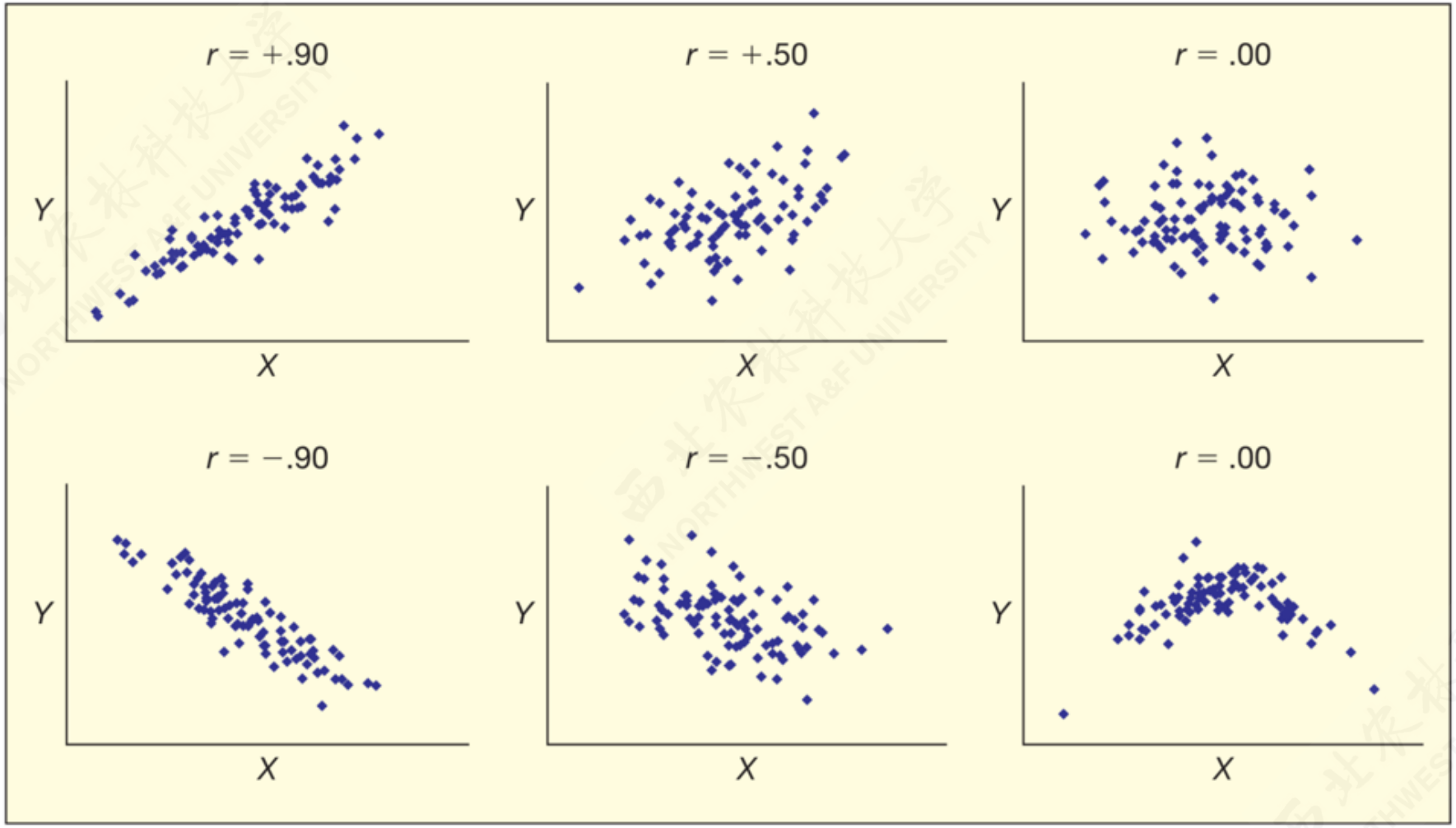
相关关系的描述与测度：散点图



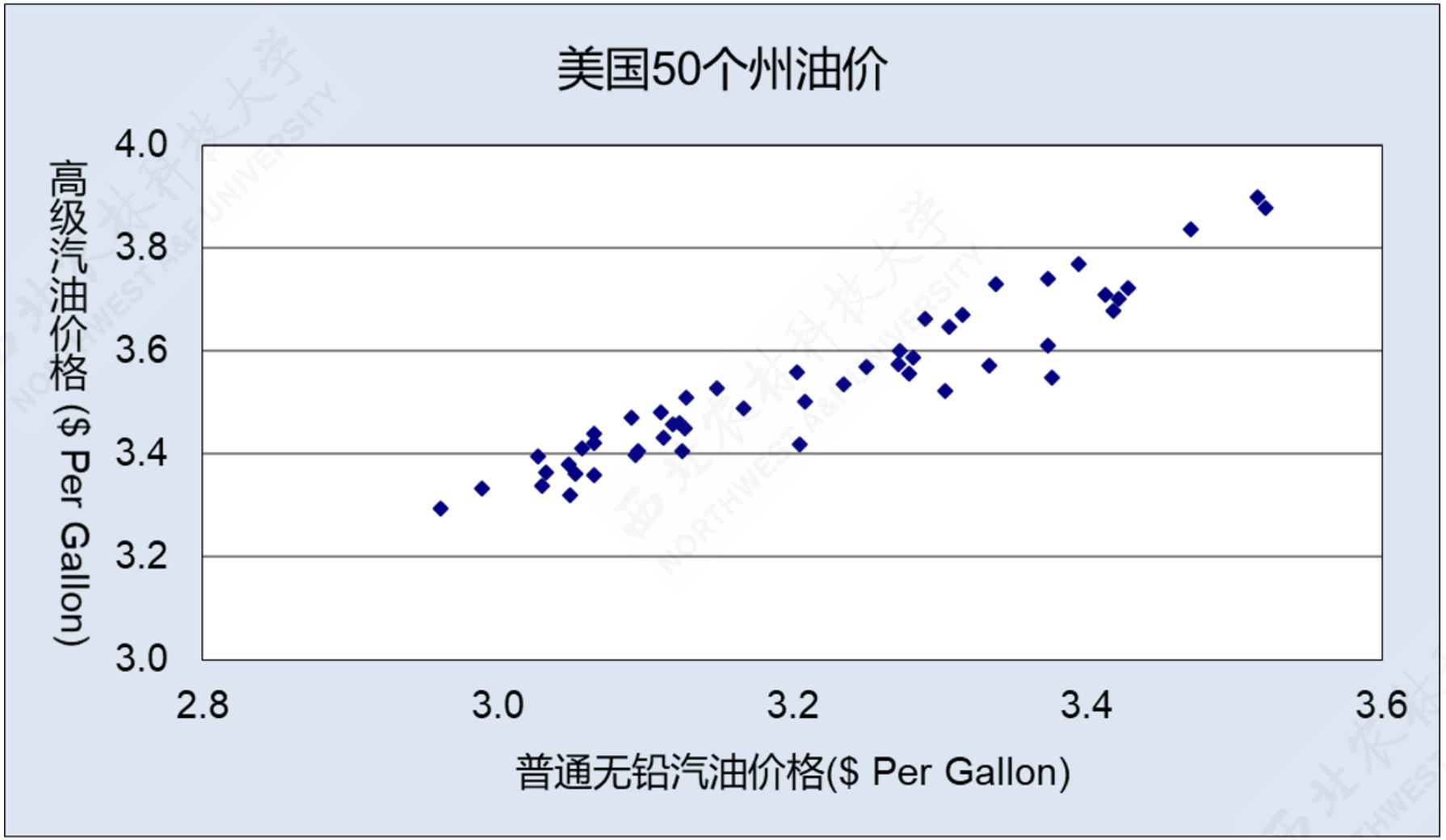
相关关系的描述与测度：散点图



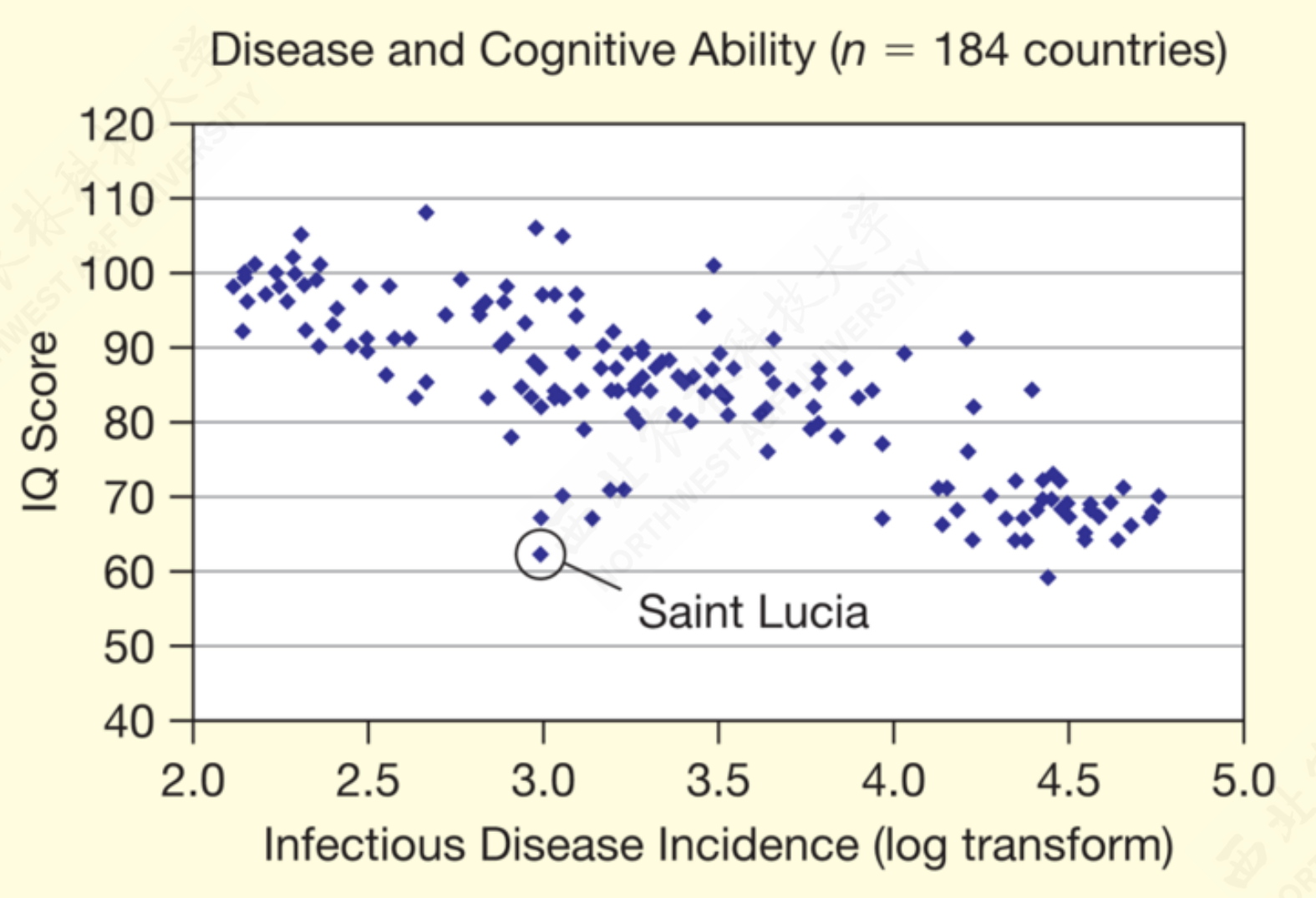
相关关系的描述与测度：散点图



（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：

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

简单相关系数的小ff计算公式：

`$$\begin{align} r & = \frac{ \sum{\left((Xi - \overline{X})(Y_i - \overline{Y})\right) } }{\sqrt{\sum{(X_i - \overline{X})^2 }\sum{(Y_i - \overline{Y})^2}}} = \frac{S S{XY}}{\sqrt{S S{XX}} \sqrt{S S{YY}}} = \frac{\sum{x_iy_i}}{\sqrt{\sum{x_i^2}\sum{y_i^2}}}

\end{align}$$`

其中：

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。

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：

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

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

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。

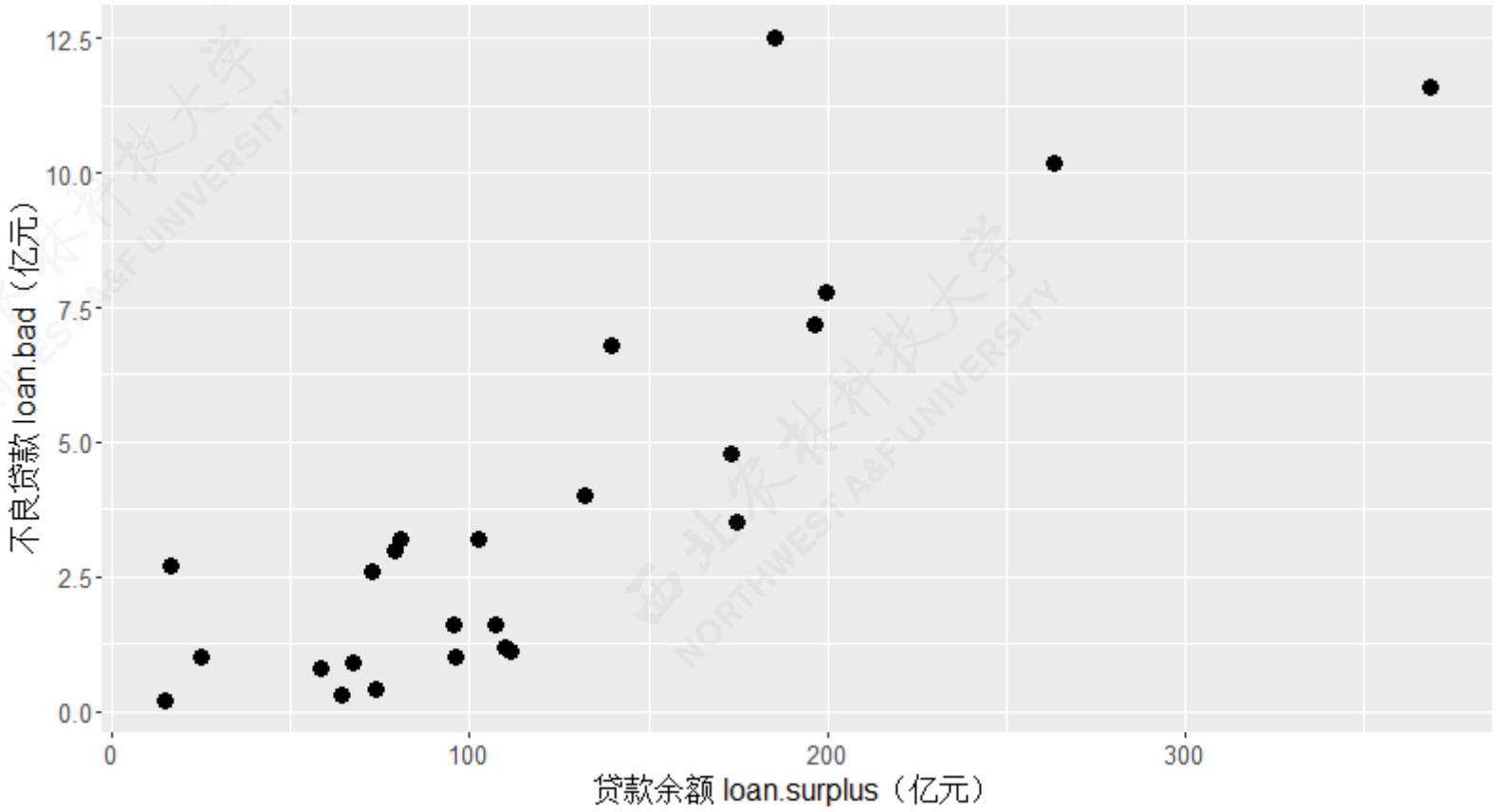
ID.bank	loan.bad	loan.surplus	loan.receivable	loan.numbers	investment.fixe
1	0.9	67.3	6.8	5	51.9
2	1.1	111.3	19.8	16	90.9
3	4.8	173	7.7	17	73.7
4	3.2	80.8	7.2	10	14.5
5	7.8	199.7	16.5	19	63.2
6	2.7	16.2	2.2	1	2.2
7	1.6	107.4	10.7	17	20.2

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）

1）大数计算表

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（大数）

1）大数计算表

2）r计算式1

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

（案例）银行贷款：不良贷款VS贷款余额的相关系数（小的）

1）小的计算表

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

（案例）银行贷款：不良贷款VS贷款余额的相关系数

1）小的计算表

2）r计算式2

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

（案例）银行贷款：相关系数矩阵(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
```

5.2 回归分析的基本思想

相关关系VS因果关系

重要概念

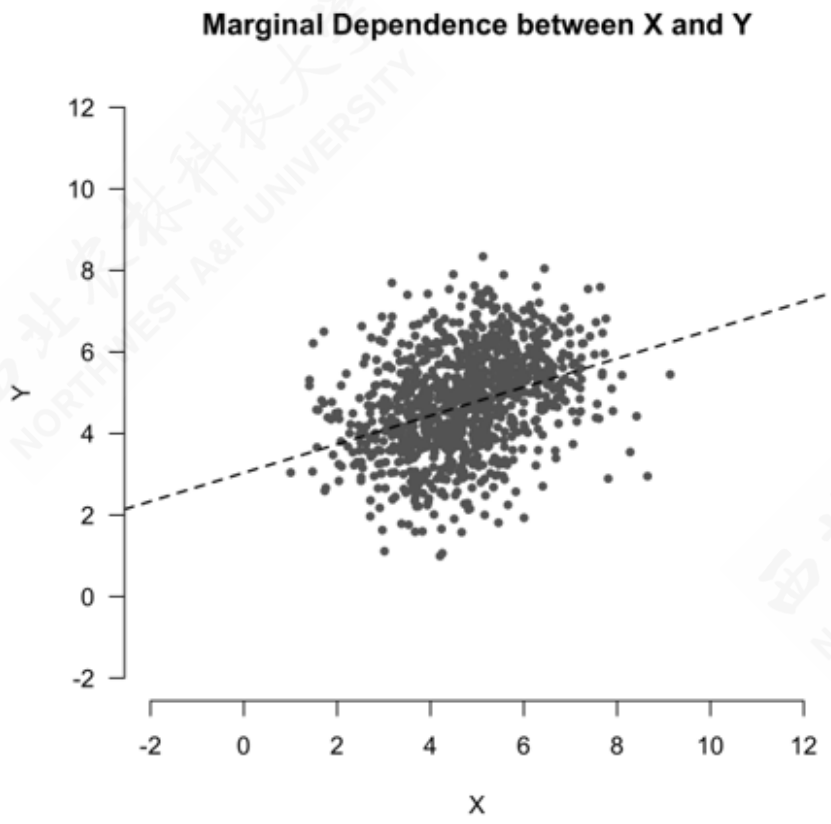
线性回归分析

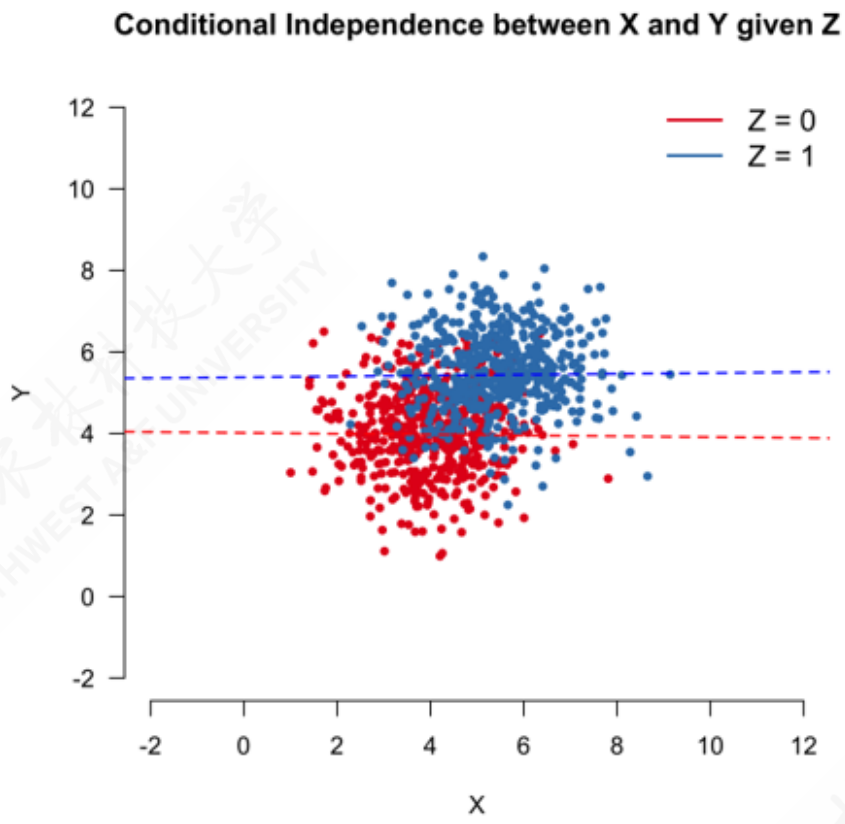
从一组样本数据出发，确定变量之间的数学关系式。

对这些关系式的可信程度进行各种统计检验，并从影响某一特定变量的诸多变量中找出哪些变量的影响显著，哪些不显著。

利用所求的关系式，根据一个或几个变量的取值来预测或控制另一个特定变量的取值，并给出这种预测或控制的精确程度。

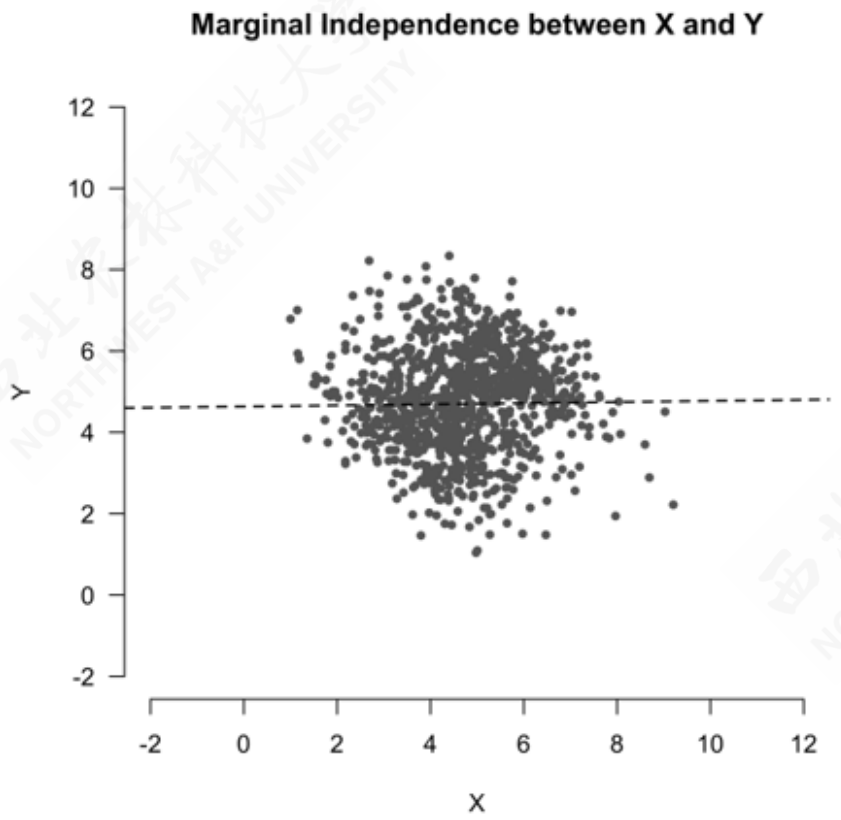
相关关系：边际相关与条件相关I

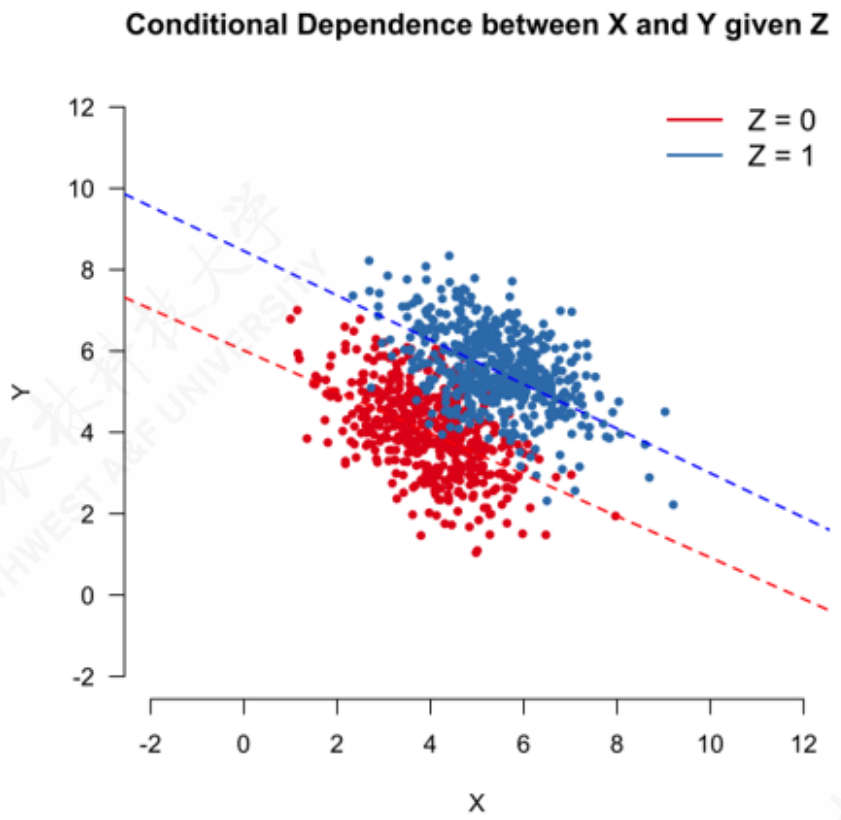




边际相关但是条件独立

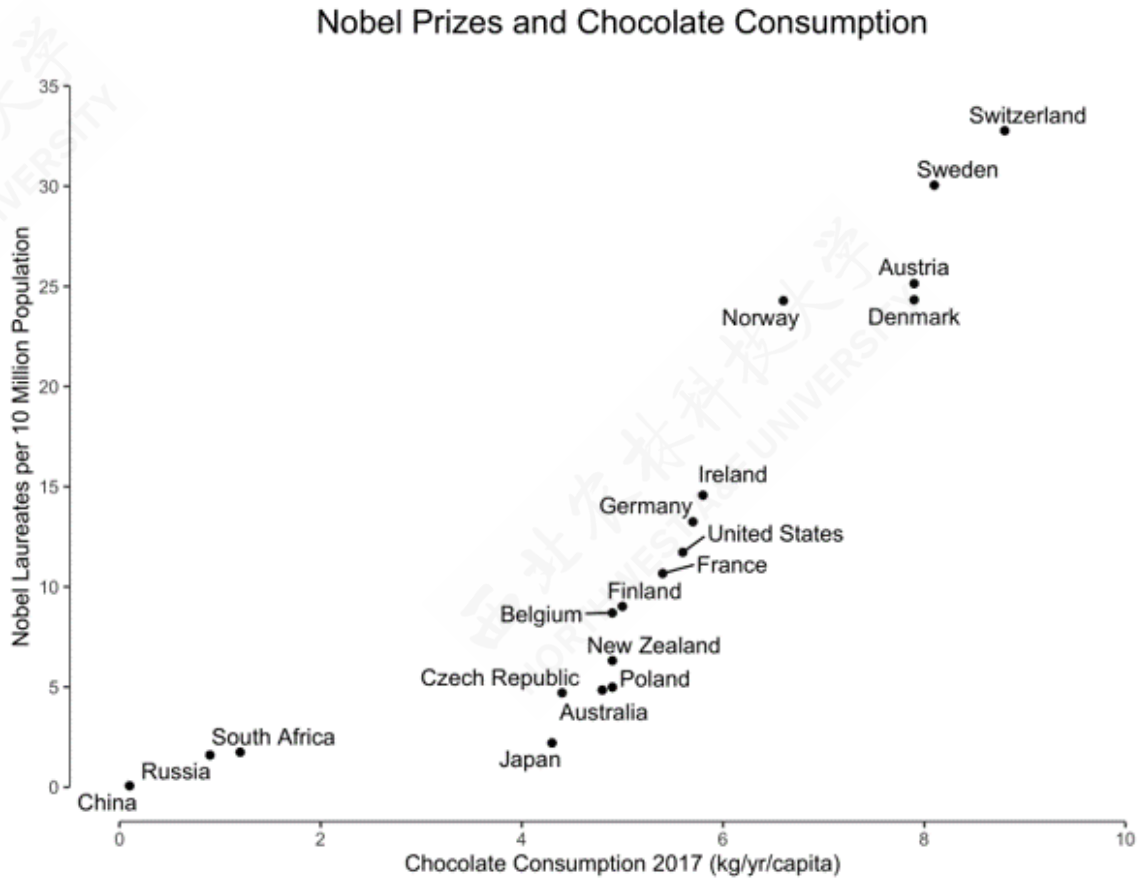
相关关系：边际相关与条件相关2





边际独立但是条件相关

相关关系VS因果关系

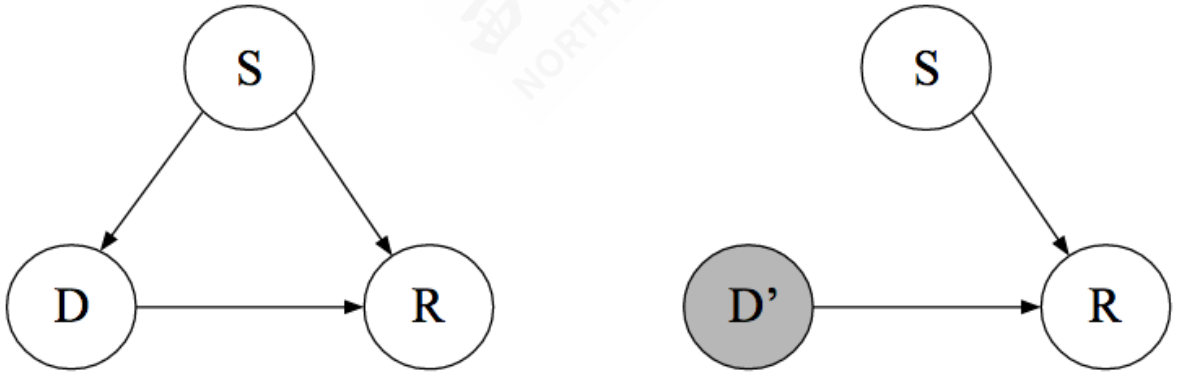


巧克力消费量与诺贝尔奖数量

相关关系VS因果关系：性别的作用

	Drug	No Drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Men & Women	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

治疗康复表

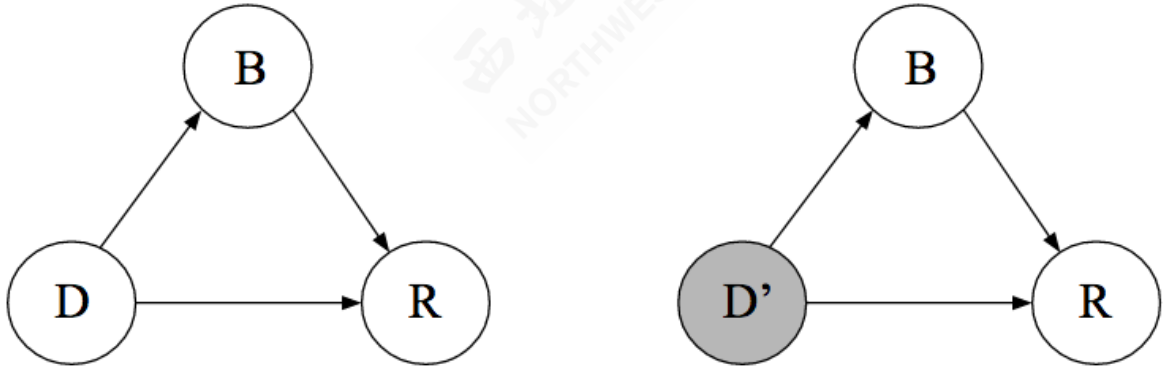


因果关系图

相关关系VS因果关系：血压的作用

	Drug	No Drug
Low Blood pressure	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
High Blood pressure	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Low & High Blood pressure	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

治疗康复表



因果关系图

（案例）微型家庭总体

（案例）假想总体：60个家庭的收支数据（直观列表）

X，每周家庭收入（美元）										
Y \ X	80	100	120	140	160	180	200	220	240	260
Y，每周家庭消费支出	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
	—	88	—	113	125	140	—	160	189	185
	—	—	—	115	—	—	—	162	—	191
小计	325	462	445	707	678	750	685	1043	966	1211
合计	7272									

60个家庭的收入和支出情况：假设的总体

（案例）假想总体：60个家庭的收支数据（扁数据形态）

60个家庭的收入和支出情况：假设的总体

Mark	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
X	80	100	120	140	160	180	200	220	240	260
Y1	55	65	79	80	102	110	120	135	137	150
Y2	60	70	84	93	107	115	136	137	145	152
Y3	65	74	90	95	110	120	140	140	155	175
Y4	70	80	94	103	116	130	144	152	165	178
Y5	75	85	98	108	118	135	145	157	175	180
Y6		88		113	125	140		160	189	185
Y7				115				162		191

（案例）假想总体：60个家庭的收支数据（长数据形态）

60个家庭的收入和支出情况：假设的总体

id	group	X	Y
1	1	80	55
2	1	80	60
3	1	80	65
4	1	80	70
5	1	80	75
6	2	100	65
7	2	100	70
8	2	100	74

Showing 1 to 8 of 60 entries

Previous 1 2 3 4 5 … 8 Next

重要概念：无条件概率和无条件期望

无条件概率：

- 定义：不受 X_i变量取值影响下， Y_i出现的可能性。
- 记号：离散变量 $P(Y_i)$；连续变量 $g(Y)$

无条件期望：

- 定义：不受 X_i变量取值影响下，变量 Y_i的期望值。
- 记号： $g(Y_i)$表示连续变量的概率密度函数（cdf）

$$E(Y) = \sum_{1}^{N} Y_i \cdot P(Y_i) \qquad \text{(discrete vars)}$$

$$E(Y) = \int Y_i \cdot g(Y_i) dY \qquad \text{(continue vars)}$$

（示例）无条件概率和无条件期望的示例计算

	X，每周家庭收入（美元）									
	80	100	120	140	160	180	200	220	240	260
Y，每周家庭消费支出	55 1/60	65 1/60	79 1/60	80 1/60	102 1/60	110 1/60	120 1/60	135 1/60	137 1/60	150 1/60
	60 1/60	70 1/60	84 1/60	93 1/60	107 1/60	115 1/60	136 1/60	137 1/60	145 1/60	152 1/60
	65 1/60	74 1/60	90 1/60	95 1/60	110 1/60	120 1/60	140 1/60	140 1/60	155 1/60	175 1/60
	70 1/60	80 1/60	94 1/60	103 1/60	116 1/60	130 1/60	144 1/60	152 1/60	165 1/60	178 1/60
	75 1/60	85 1/60	98 1/60	108 1/60	118 1/60	135 1/60	145 1/60	157 1/60	175 1/60	180 1/60
	— —	88 1/60	— —	113 1/60	125 1/60	140 1/60	— —	160 1/60	189 1/60	185 1/60
	— —	— —	— —	115 1/60	— —	— —	— —	162 1/60	— —	191 1/60
小计	325 —	462 —	445 —	708 —	678 —	750 —	685 —	1043 —	966 —	1211 —
无条件期望										

无条件概率和无条件期望

（示例）无条件期望的计算过程

$$
\begin{aligned}
E(Y) &= \sum_{1}^{N} Y_i \cdot P(Y_i) \\
&= \sum_{1}^{60} \left(55 * \frac{1}{60} + 60 * \frac{1}{60} + \cdots + 191 * \frac{1}{60} \right) \\
&= \frac{1}{60} \sum_{1}^{60} Y_i \\
&= \frac{7272}{60} \\
&= 121.2
\end{aligned}
$$

重要概念：条件概率和条件期望

条件概率：

- 定义：给定变量 X_i的取值条件下， Y_i出现的可能性。
- 记号：离散变量 $P(Y_i|X_i)$；连续变量 $g(Y|X)$

条件期望：

- 在给定变量 X_i的取值条件下， Y_i的期望值。
- 记号： $g(Y|X)$表示连续变量的条件概率密度函数（cdf）

$$E(Y|X_i) = \sum_{1}^{N} (Y_i|X_i) \cdot P(Y_i|X_i) \qquad \text{(discrete vars)}$$

$$E(Y|X_i) = \int (Y|X) \cdot g(Y|X) dY \qquad \text{(continue vars)}$$

（示例）条件概率和条件期望的计算

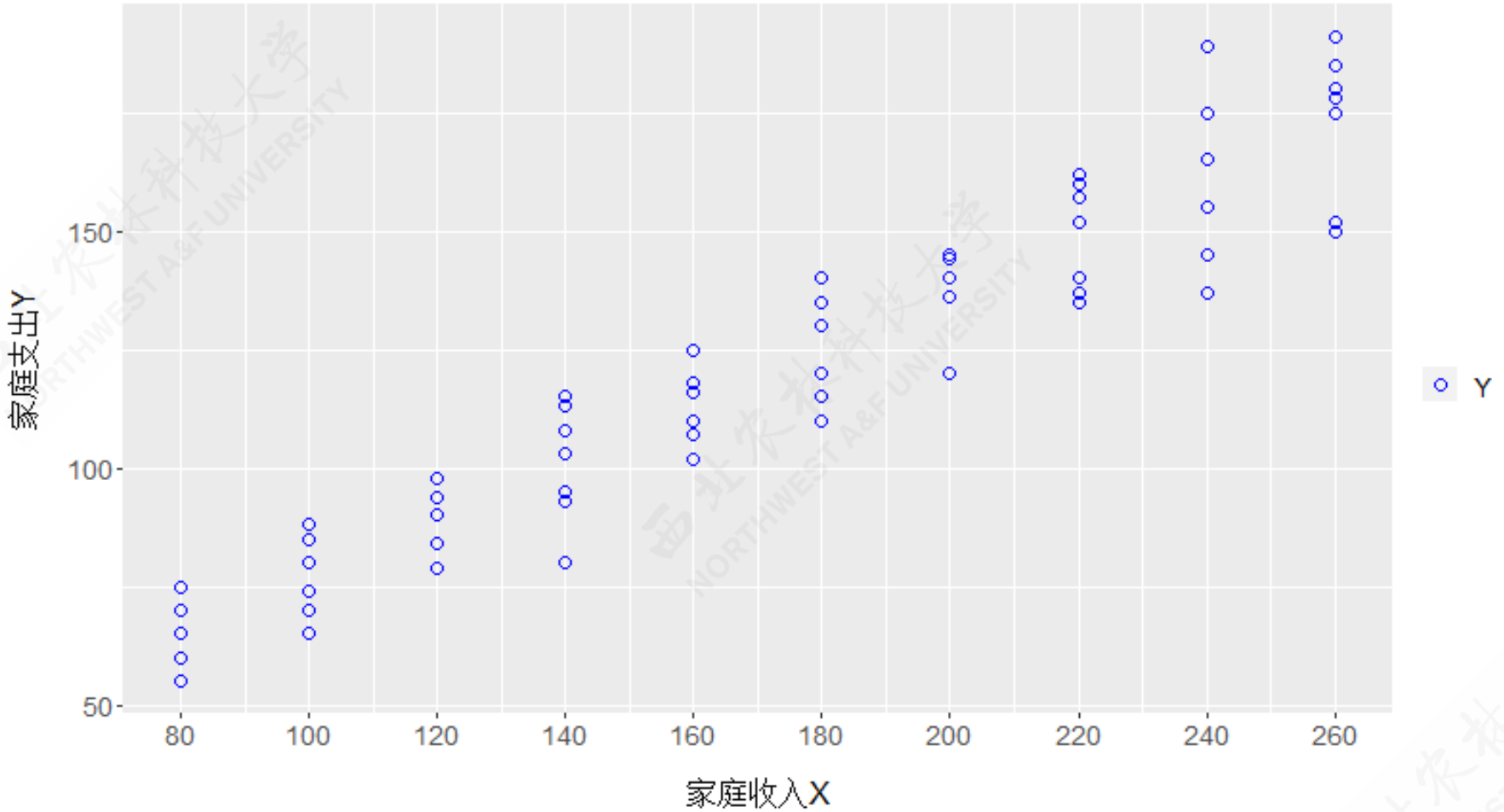
	X，每周家庭收入（美元）																					
	80		100		120		140		160		180		200		220		240		260			
Y，每周家庭消费支出	55	1/5	65	1/6	79	1/5	80	1/7	102	1/6	110	1/6	120	1/5	135	1/7	137	1/6	150	1/7		
	60	1/5	70	1/6	84	1/5	93	1/7	107	1/6	115	1/6	136	1/5	137	1/7	145	1/6	152	1/7		
	65	1/5	74	1/6	90	1/5	95	1/7	110	1/6	120	1/6	140	1/5	140	1/7	155	1/6	175	1/7		
	70	1/5	80	1/6	94	1/5	103	1/7	116	1/6	130	1/6	144	1/5	152	1/7	165	1/6	178	1/7		
	75	1/5	85	1/6	98	1/5	108	1/7	118	1/6	135	1/6	145	1/5	157	1/7	175	1/6	180	1/7		
	—	—	88	1/6	—	—	113	1/7	125	1/6	140	1/6	—	—	160	1/7	189	1/6	185	1/7		
	—	—	—	—	—	—	115	1/7	—	—	—	—	—	—	162	1/7	—	—	191	1/7		
小计	325	1	462	1	445	1	708	1	678	1	750	1	685	—	1043	1	966	—	1211	1		
条件期望	65		77		89		101		113		125		137		149		161		173			

条件概率和条件期望

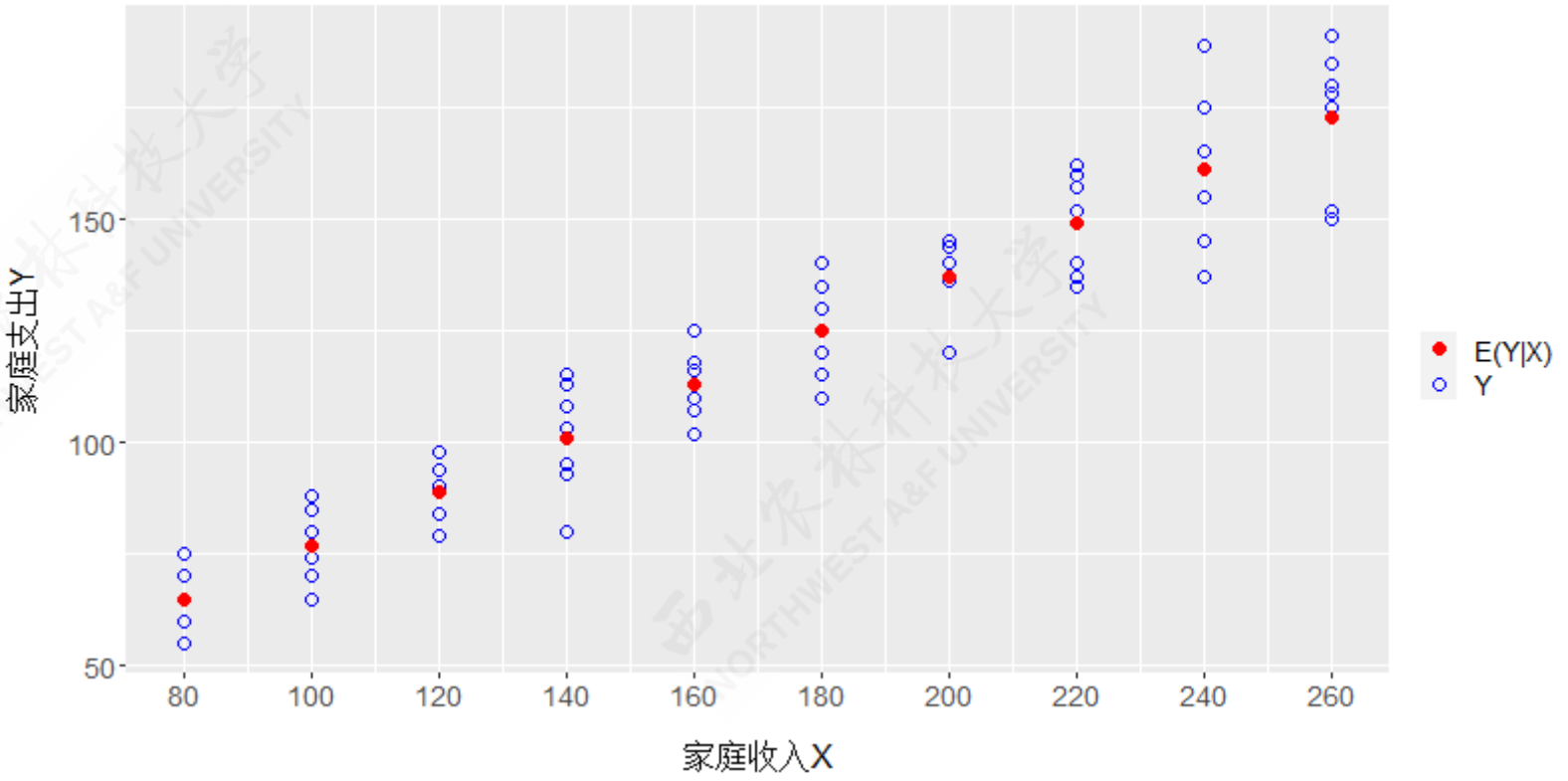
（示例）条件期望的计算过程

$$
\begin{aligned}
E(Y|80) &= \sum_{1}^{N} Y_i \cdot P(Y_i|X=80) \\
&= \sum_{1}^{5} \left(55 * \frac{1}{5} + 60 * \frac{1}{5} + \cdots + 75 * \frac{1}{5} \right) \\
&= \frac{1}{5} \sum_{1}^{5} Y_i \\
&= \frac{325}{5} \\
&= 65
\end{aligned}
$$

（示例）假想总体的全部数据展示

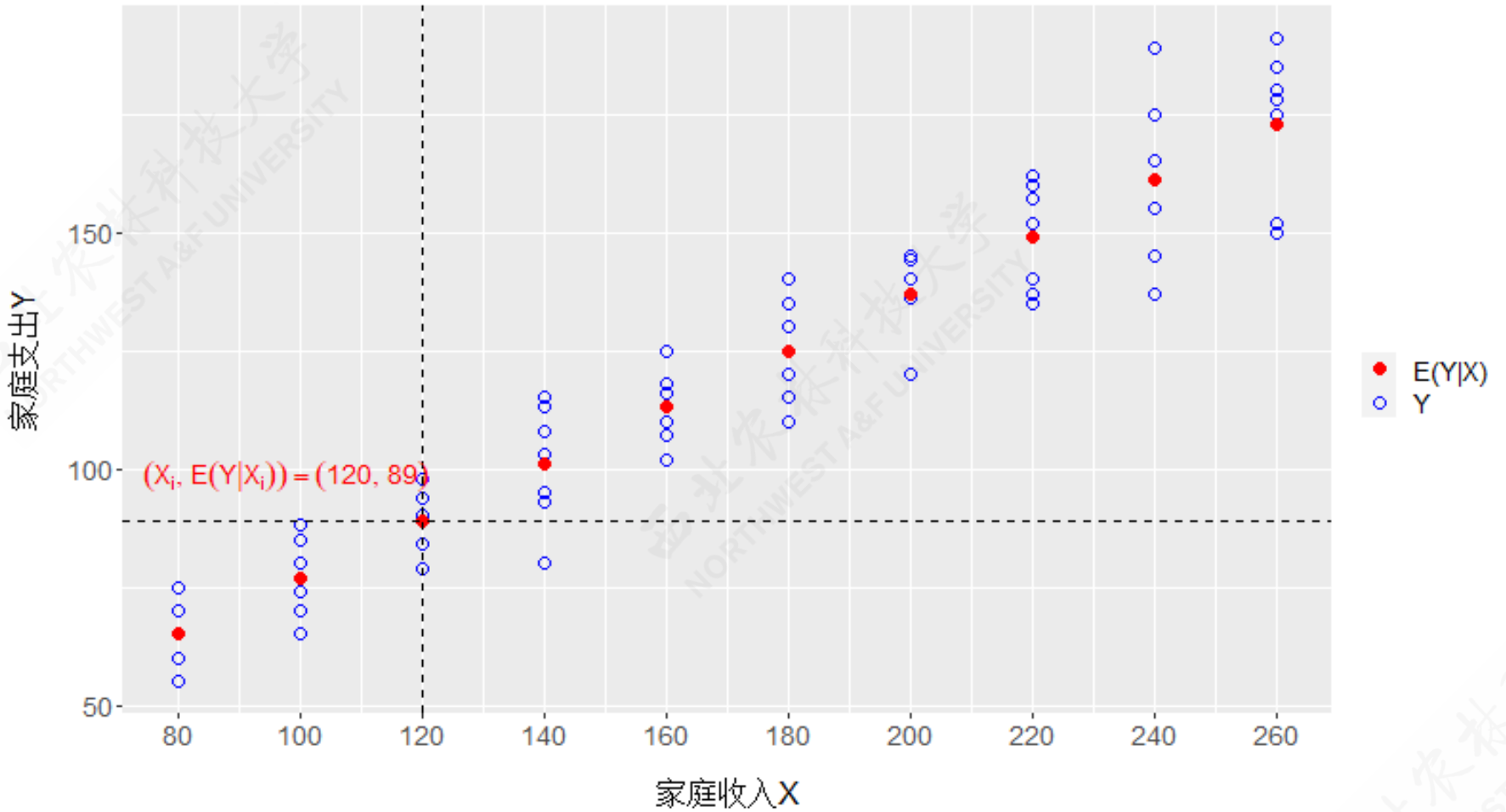


（示例）给定不同X水平下Y条件期望值



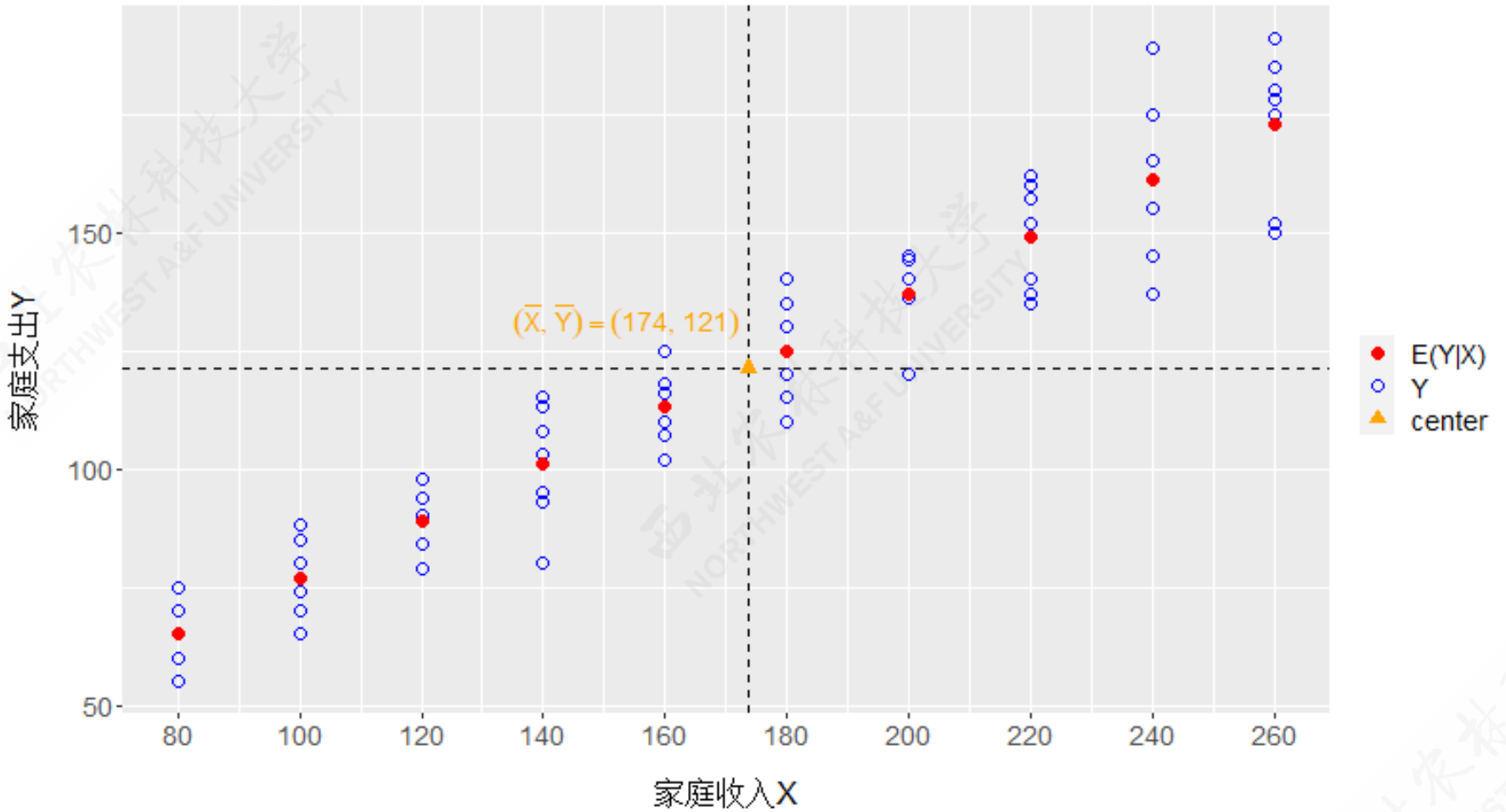
var	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
X	80	100	120	140	160	180	200	220	240	260
E(Y|X)	65	77	89	101	113	125	137	149	161	173

（示例）给定不同X水平下Y条件期望值



给定 $X=120$水平下 Y条件期望值 $E(Y|X_i=120)=89$

（示例）X均值和Y的无条件期望值



X的均值 $\bar{X}$ =173.67和Y的无条件期望值 $E(Y) = 121.20$

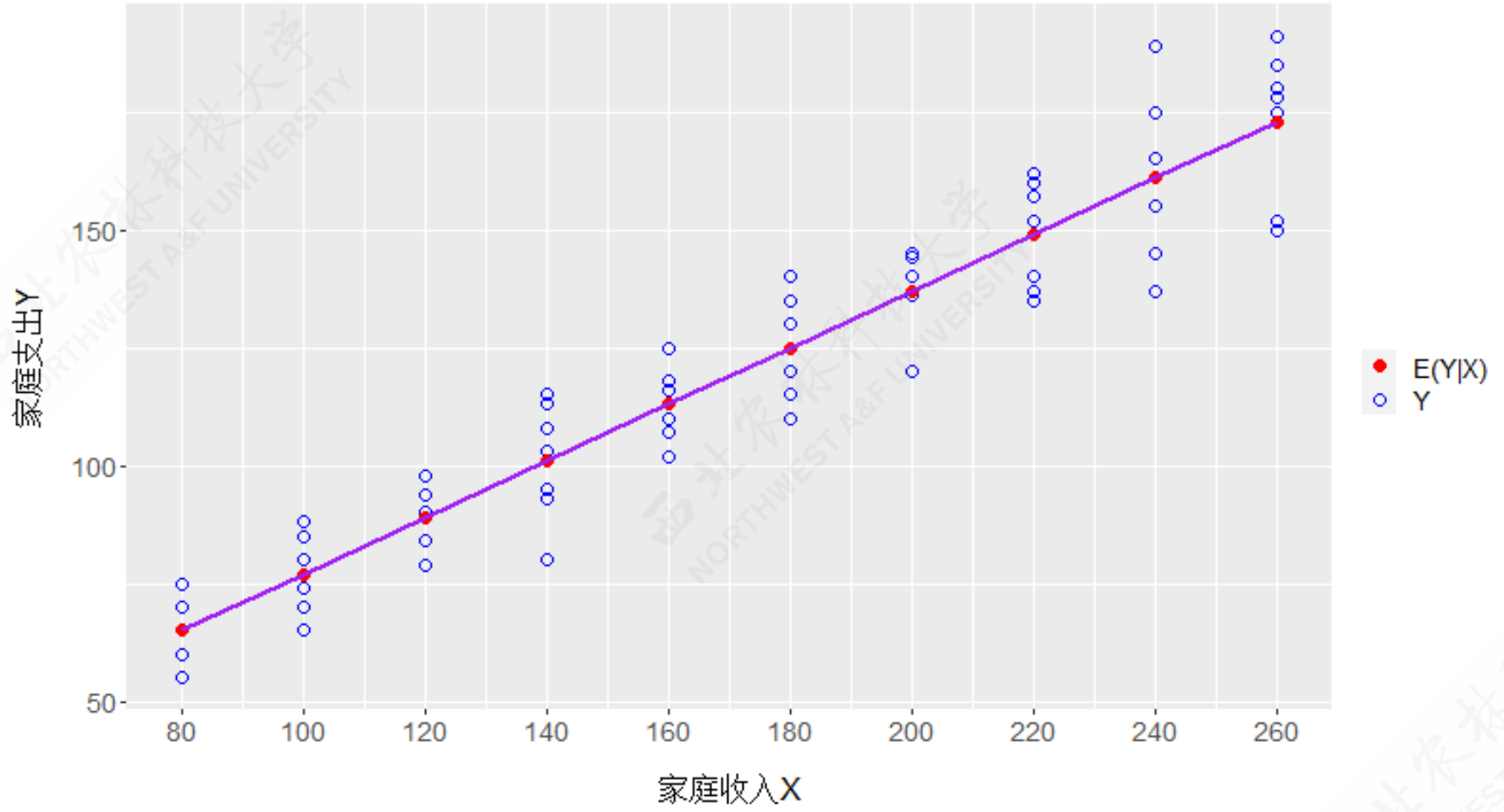
重要概念：总体回归线（PRL）

- 几何：给定X值时Y的条件期望值的轨迹。
- 统计：实质上就是Y对X的回归。

总体回归曲线(Population Regression Curve，PRC)：条件期望值的轨迹表现为一条曲线(Curve)。

总体回归线(Population Regression Line，PRL)：条件期望值的轨迹表现为一条直线(Line)。

重要概念：总体回归线（PRL）



总体回归线PRL

重要概念：总体回归函数（PRF）

总体回归函数（Population Regression Function，PRF）：它是对总体回归曲线(PRC)的数学函数表现形式。

如果不知道总体回归曲线的具体形式，则总体回归函数PRF表达为如下隐函数形式（PRF）：

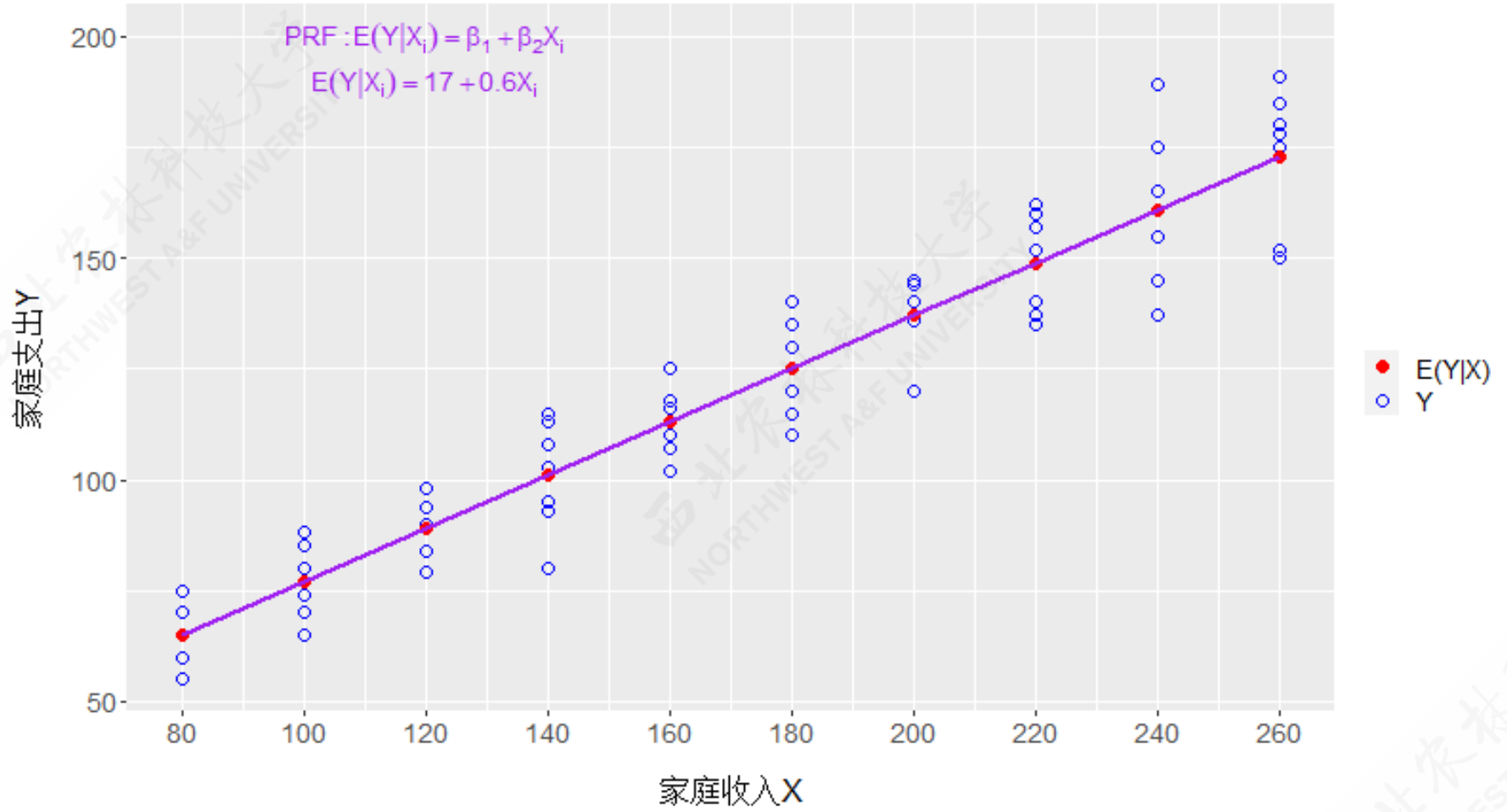
$$E(Y|X_i) = f(X_i) \qquad \text{(PRF)}$$

如果总体回归曲线是直线形式，则总体回归函数PRF表达为如下显函数形式（PRF_L）：

$$E(Y|X_i) = \beta_1 + \beta_2 X_i \qquad \text{(PRF\_L)}$$

- β_1, β_2分别称为截距(intercept)和斜率系数(slope coefficient)。
- β_1, β_2称为总体参数或回归系数(regression coefficients)。
- β_1, β_2为未知但却是固定的参数。

重要概念：总体回归函数（PRF）



总体回归线PRL与总体回归函数PRF

重要概念：总体回归模型（PRM）

总体回归模型（Population Regression model, PRM）：把总体回归函数表达成**随机设定**形式。

如果总体回归函数为隐函数，则**总体回归模型**记为：

$$\begin{aligned} Y_i &= E(Y|X_i) + u_i \\ &= f(X_i) + u_i \end{aligned}$$

如果总体回归函数为线性函数，则**总体回归模型**记为：

$$\begin{aligned} Y_i &= E(Y|X_i) + u_i \\ &= \beta_1 + \beta_2 X_i + u_i \end{aligned}$$

- 总体回归模型（PRM）属于**计量经济学模型**，而总体回归函数（PRF）是**数量经济学模型**（或数学模型）。
- 总体回归模型（PRM）能充分表达的是现实世界中 Y_i 变量的行为特征。

重要概念：随机干扰项

总体回归模型（PRM）设定下，Y_i将由两个部分组成。

- 特定家庭的支出（Y_i）= 系统性部分（$E(Y|X_i)$ + 随机部分（u_i）
- 特定家庭的支出（Y_i）= 系统性部分（$\beta_1 + \beta_2 X_i$）+ 随机部分（u_i）

随机干扰项：

- 也被称为随机误差项(stochastic error term)：总体回归函数中忽略掉的但又影响着Y的全部变量的替代物，它是Y_i与条件期望（$E(Y|X_i)$）的离差。

$$u_i = Y_i - E(Y|X_i)$$

重要概念：随机干扰项

随机干扰项的来源：

- 理论的含糊：除了主变量之外，还有其它变量的影响，但不清楚，只能用μ_i代替它们。（家庭收入以外？）
- 数据的不充分：可能知道被忽略的变量，但不能得到这些变量的数量信息。（如家庭财富数据不可得）
- 核心变量与其它变量：其它变量全部或其中一些合起来影响还是很小的。（如子女、教育、性别、宗教等）
- 人类行为的内在随机性。（客观存在、固有的）
- 变量被“移花接木”而产生测量误差（如弗里德曼的持久收入和消费）
- 节省原则：为了保持一个尽可能简单的回归模型
- 错误的函数形式：有时根据数据及经验无法确定一个正确的函数形式（多元回归尤其如此）

重要概念：随机干扰项

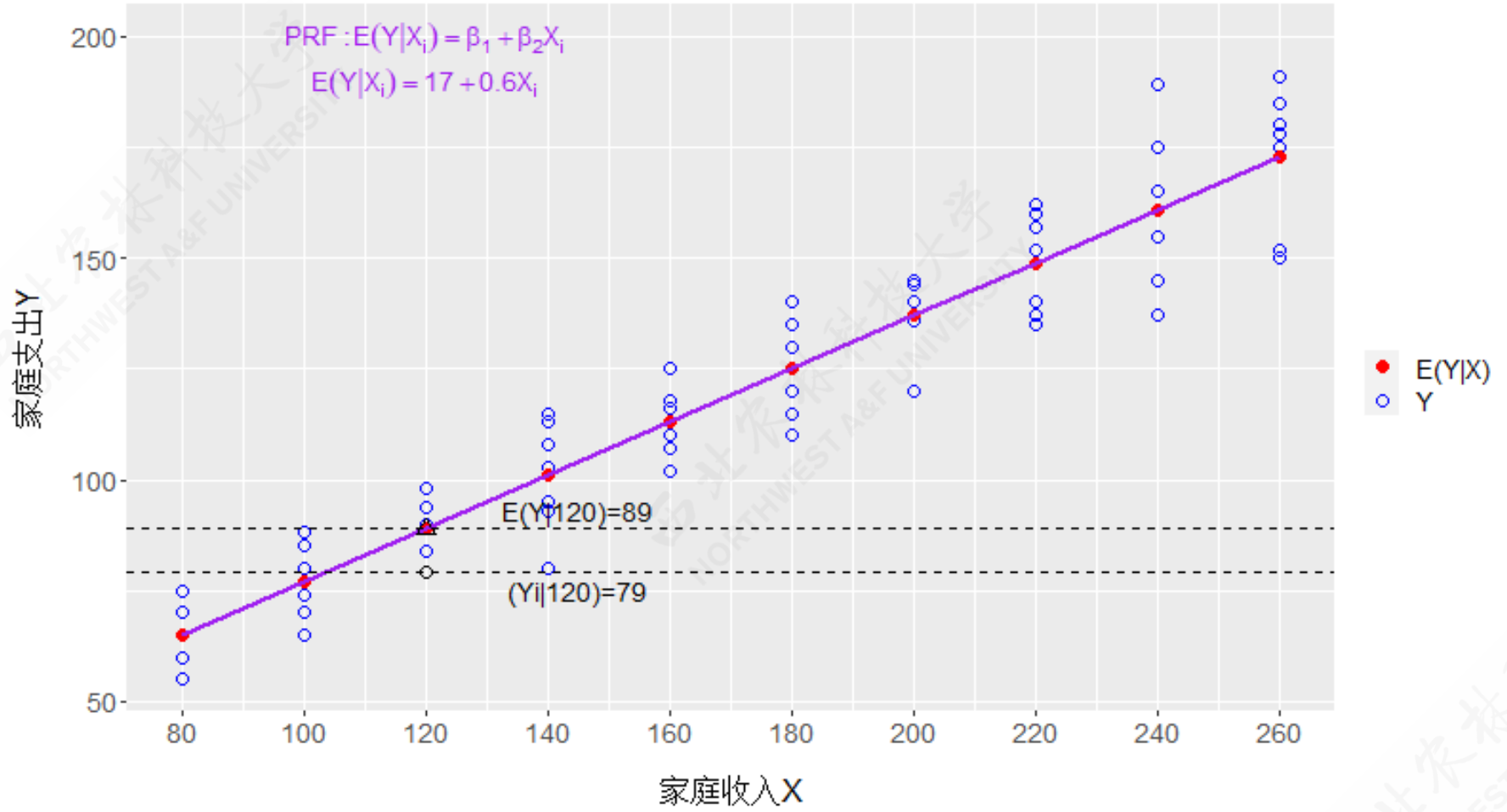
为何是“随机的”?

- 测不准？（误差）
- 测错了？（误导）
- 免不了!（内在性）

拥抱随机世界

- 风筝： Y_i
- 风筝线： $E(Y|X_i)$
- 风： u_i

重要概念：理解PRM和PRF的关系



若给定一个特定家庭 $(X_i = 120, Y_i = 79)$，则条件期望为 $E(Y|120) = 89$

重要概念：理解PRM和PRG的关系

若给定 $X_i = 120$，则5个家庭的真实消费支出分别为：

$$
\begin{aligned}
(Y_1|X = 120) &= 79 = \beta_1 + \beta_2 \cdot 120 + u_1 \\
(Y_2|X = 120) &= 84 = \beta_1 + \beta_2 \cdot 120 + u_2 \\
(Y_3|X = 120) &= 90 = \beta_1 + \beta_2 \cdot 120 + u_3 \\
(Y_4|X = 120) &= 94 = \beta_1 + \beta_2 \cdot 120 + u_4 \\
(Y_5|X = 120) &= 98 = \beta_1 + \beta_2 \cdot 120 + u_5
\end{aligned}
$$

重要概念：理解PRM和PRG的关系

主要结论：

- 总体期望刻画总体的“趋势”，总体回归线让“趋势”直观化。
- 个体随机性是不可避免的，总会“游离”于“趋势”之外。
- 随机干扰项 $u_i i$携带了随机个体的“游离”信息。
- 总体回归模型既“提取”了趋势和规律性，又“维系”着个体随机性，从而更好地表达了“真实世界”。

课后思考：

- 如果是无限总体，总体的规律性在理论上也是可以被严格表达出来么？
- 如果不告诉你总体，你怎么知道“触碰”到的是“真实的”趋势/规律？
- 从假想的60个家庭的微型总体中，“随便”抽取10个家庭的数据，你还能看到“直线”趋势么？

重要概念：“线性”的含义

“线性回归模型”中“线性”一词的含义

- **变量“线性”模型**：因变量对于自变量是线性的。
- **参数“线性”模型**：因变量对于参数是线性的。

（测试题）“线性”的含义

下列模型分别属于哪一类？请指出来：

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad (\text{mod1})$$

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + u_i \qquad (\text{mod2})$$

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + u_i \qquad (\text{mod3})$$

$$Y_i = \beta_1 + \beta_2 \frac{1}{X_i} + u_i \qquad (\text{mod4})$$

$$Y_i = \beta_1 + \beta_2 ln(X_i) + u_i \qquad (\text{mod5})$$

$$ln(Y_i) = \beta_1 + \beta_2 X_i + u_i \qquad (\text{mod6})$$

（测试题）“线性”的含义

下列模型分别属于哪一类？请指出来：

$$ln(Y_i) = \beta_1 - \beta_2 \frac{1}{X_i} + u_i \qquad (\text{mod}7)$$

$$ln(Y_i) = ln(\beta_1) + \beta_2 ln(X_i) + u_i \qquad (\text{mod}8)$$

$$Y_i = \frac{1}{1 + e^{(\beta_1 + \beta_2 X_{2i} + u_i)}} \qquad (\text{mod}9)$$

$$Y_i = \beta_1 + (0.75 - \beta_1) e^{-\beta_2 (X_i - 2)} + u_i \qquad (\text{mod}10)$$

$$Y_i = \beta_1 + \beta_2^3 X_i + u_i \qquad (\text{mod}11)$$

重要概念：样本回归线(SRL)

样本(Sample)：

- 从总体中随机抽取得到的数据。

样本回归线(Sample Regression Line，SRL)：

- 是通过拟合**样本数据**得到的一条曲线（或直线）。换言之，这条线由拟合值$\hat{Y}_i$连接而成。
- $\hat{Y}_i$是对条件期望值$Y|X_i$的拟合。
- 拟合方法有很多，例如采用OLS方法对样本数据进行拟合。
  - 尽可能拟合数据
  - 用什么方法拟合？
  - 曲线是什么形态？

重要概念：样本回归函数(SRF)

样本回归函数(Sample Regression Function，SRF)：是样本回归曲线的数学函数形式，可是是线性的或非线性。如果是直线则可以写成：

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

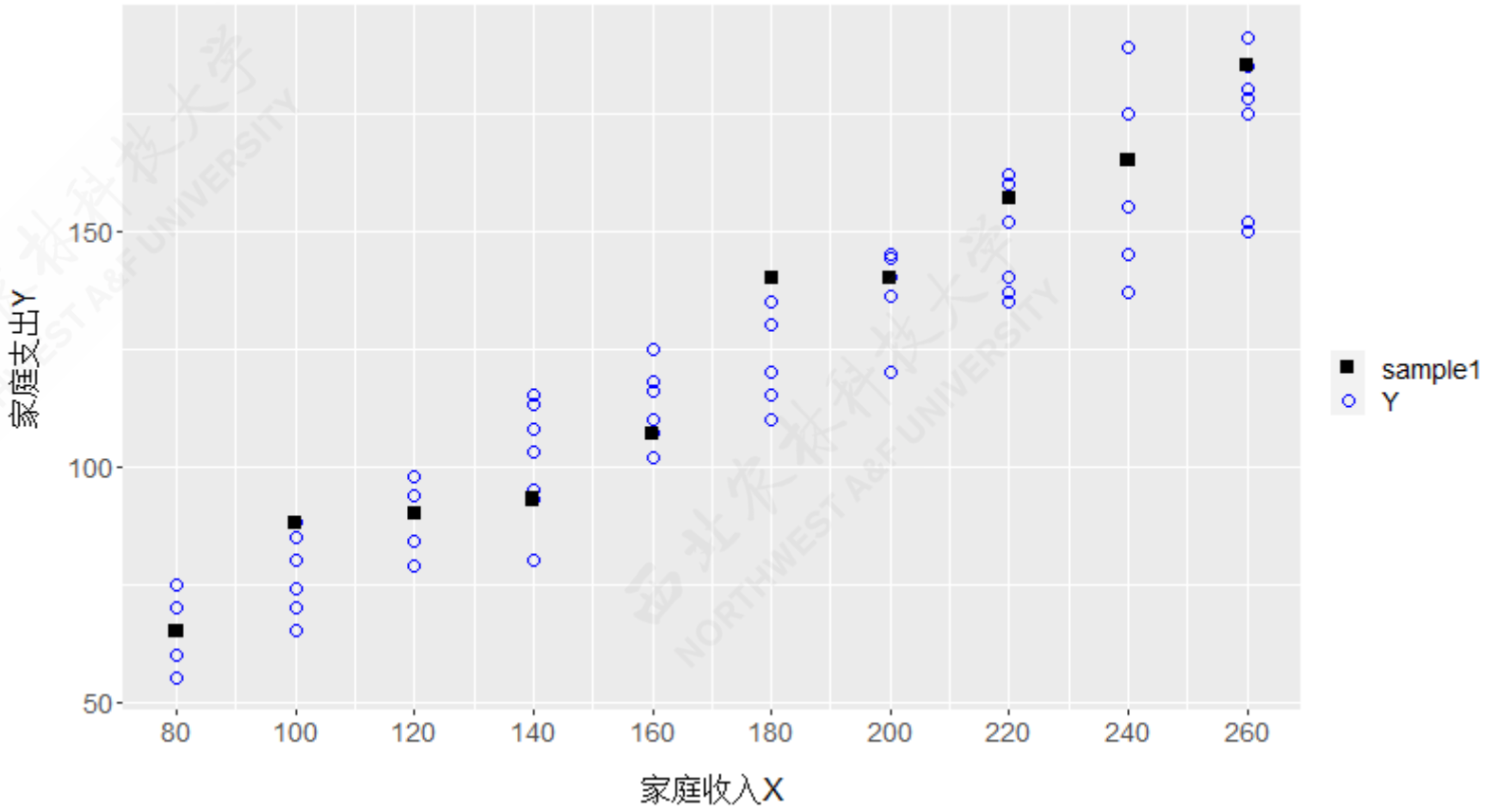
对比总体回归函数（PRF）：

$$E(Y|X_i) = \beta_1 + \beta_2 X_i$$

可以认为：

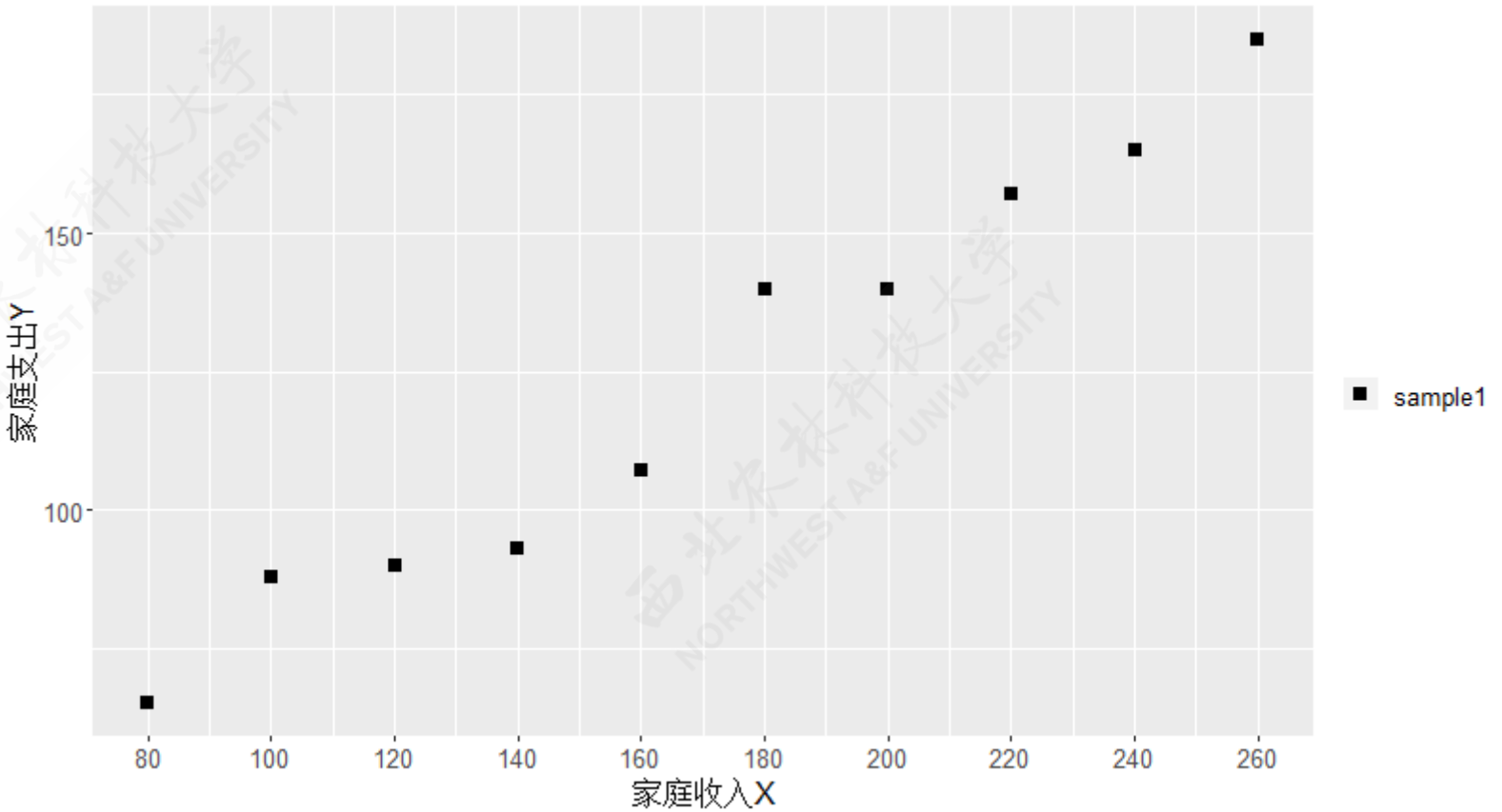
- $\hat{Y}_i$是对 $E(Y|X_i)$的估计量。
- $\hat{\beta}_1$是对 β_1的估计量。
- $\hat{\beta}_2$是对 β_2的估计量。

（示例）第一份随机样本：抽样



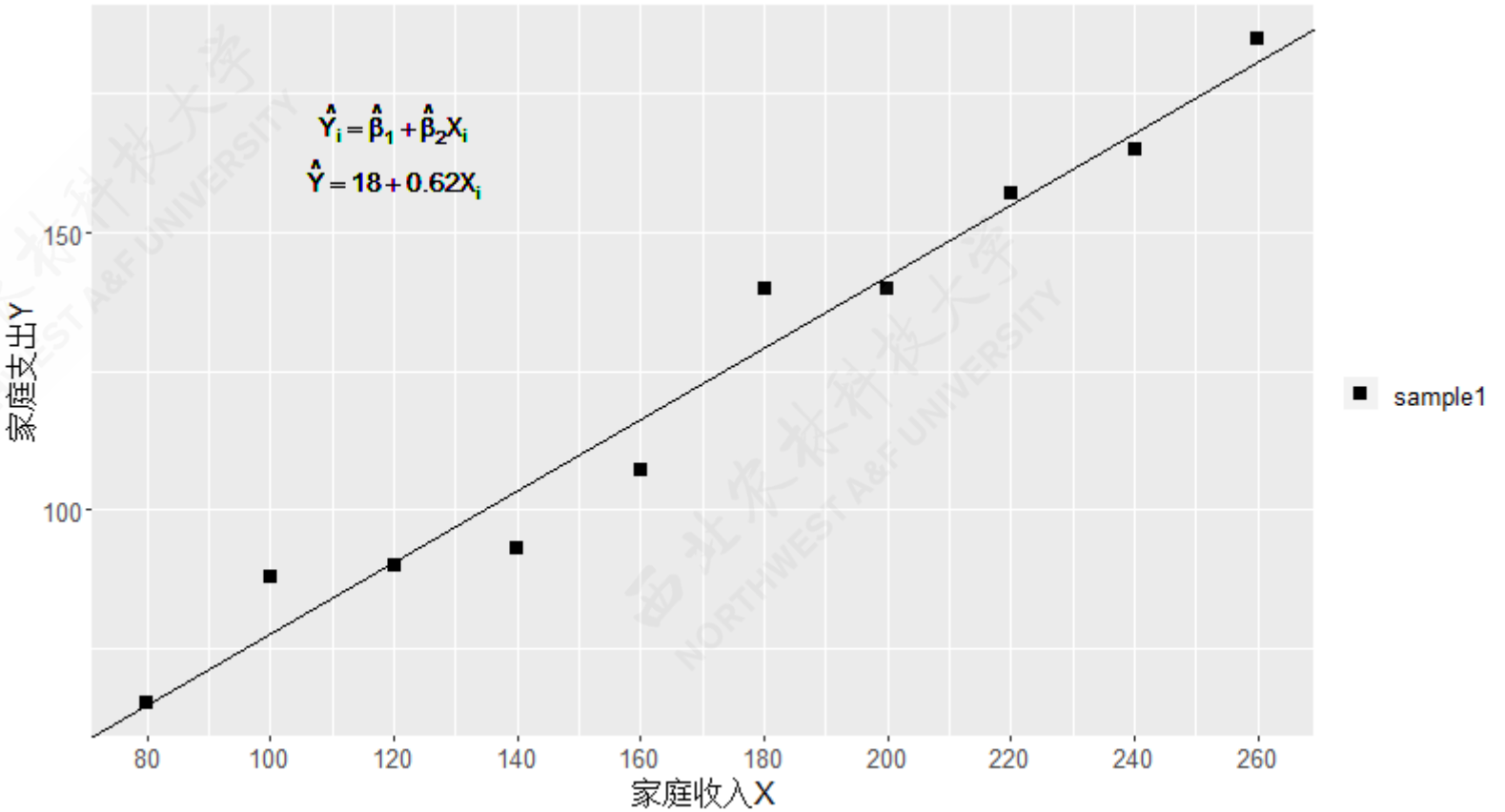
var	n1	n2	n3	n4	n5	n6	n7	n8	n9	n10
X	80	100	120	140	160	180	200	220	240	260
Y	65	88	90	93	107	140	140	157	165	185

（示例）第一份随机样本：数据



var	n1	n2	n3	n4	n5	n6	n7	n8	n9	n10
X	80	100	120	140	160	180	200	220	240	260
Y	65	88	90	93	107	140	140	157	165	185

（示例）第一份随机样本：SRL



var	n1	n2	n3	n4	n5	n6	n7	n8	n9	n10
X	80	100	120	140	160	180	200	220	240	260
Y	65	88	90	93	107	140	140	157	165	185

（示例）第一份随机样本：SRS

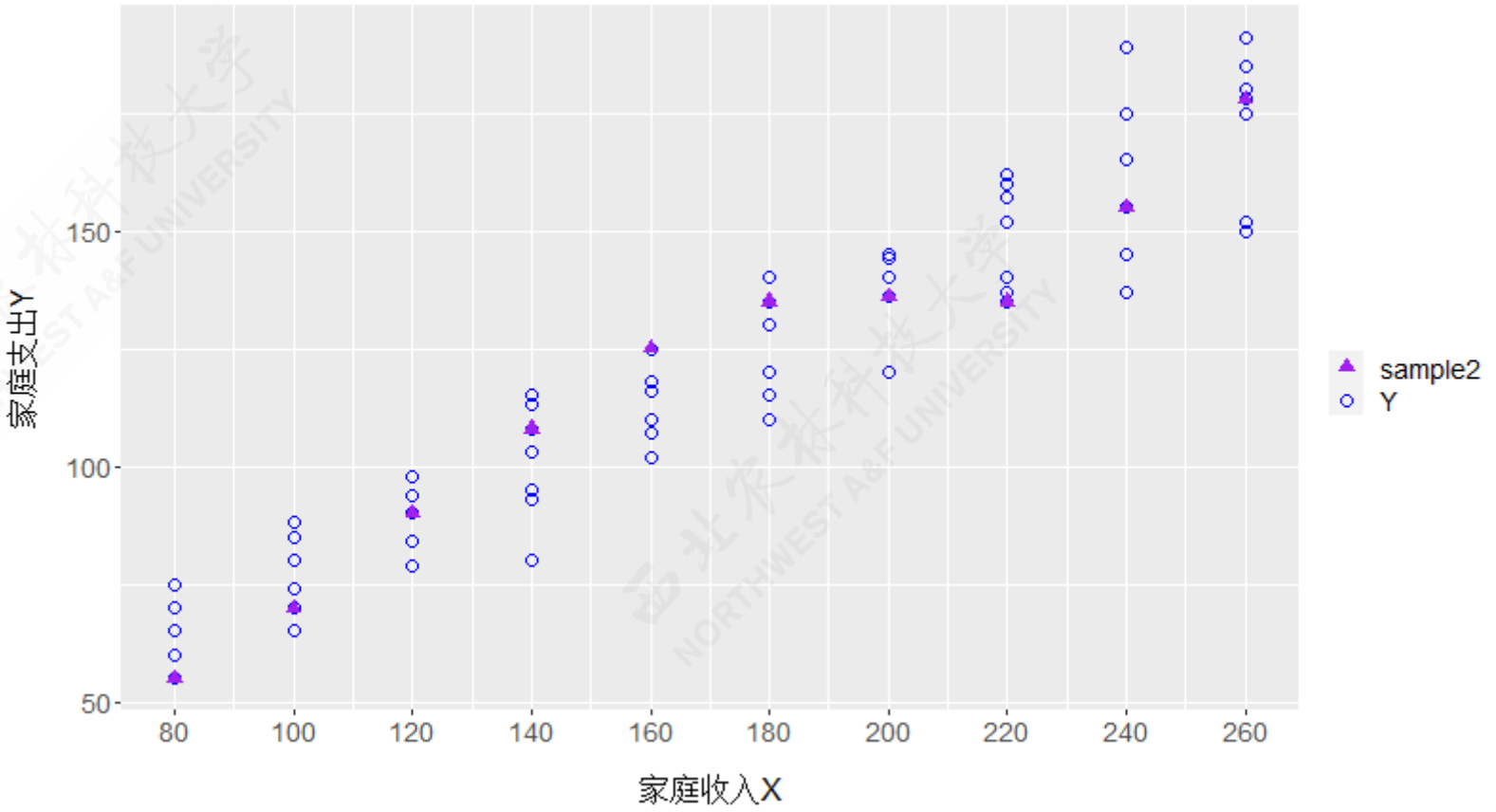
根据第一份随机样本拟合得到的**样本回归函数**SRF：

$$\hat{Y} = +13.38 + 0.64X$$

样本数据如下：

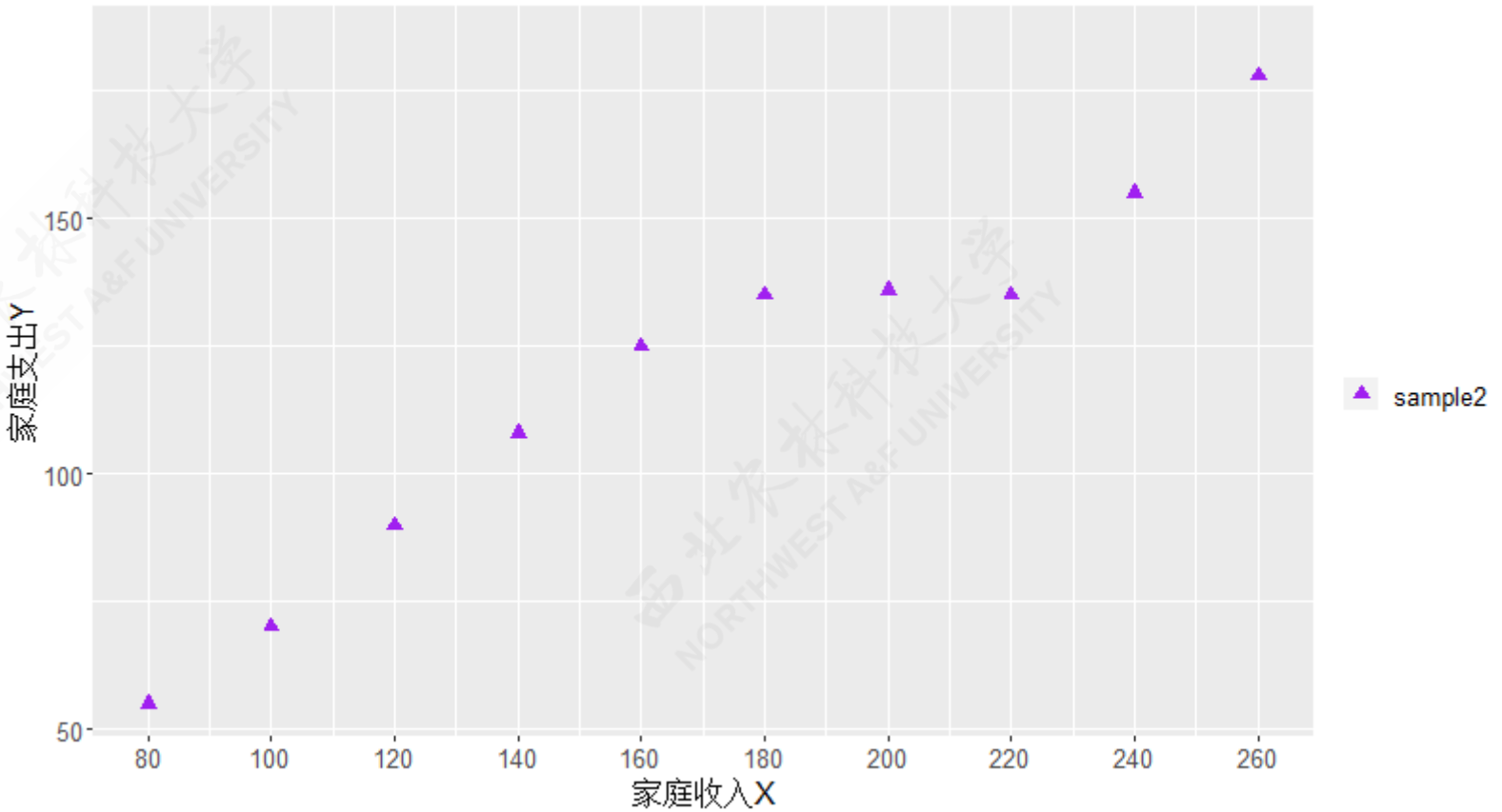
var	n1	n2	n3	n4	n5	n6	n7	n8	n9	n10
X	80	100	120	140	160	180	200	220	240	260
Y	65	88	90	93	107	140	140	157	165	185

（示例）第二份随机样本：抽样



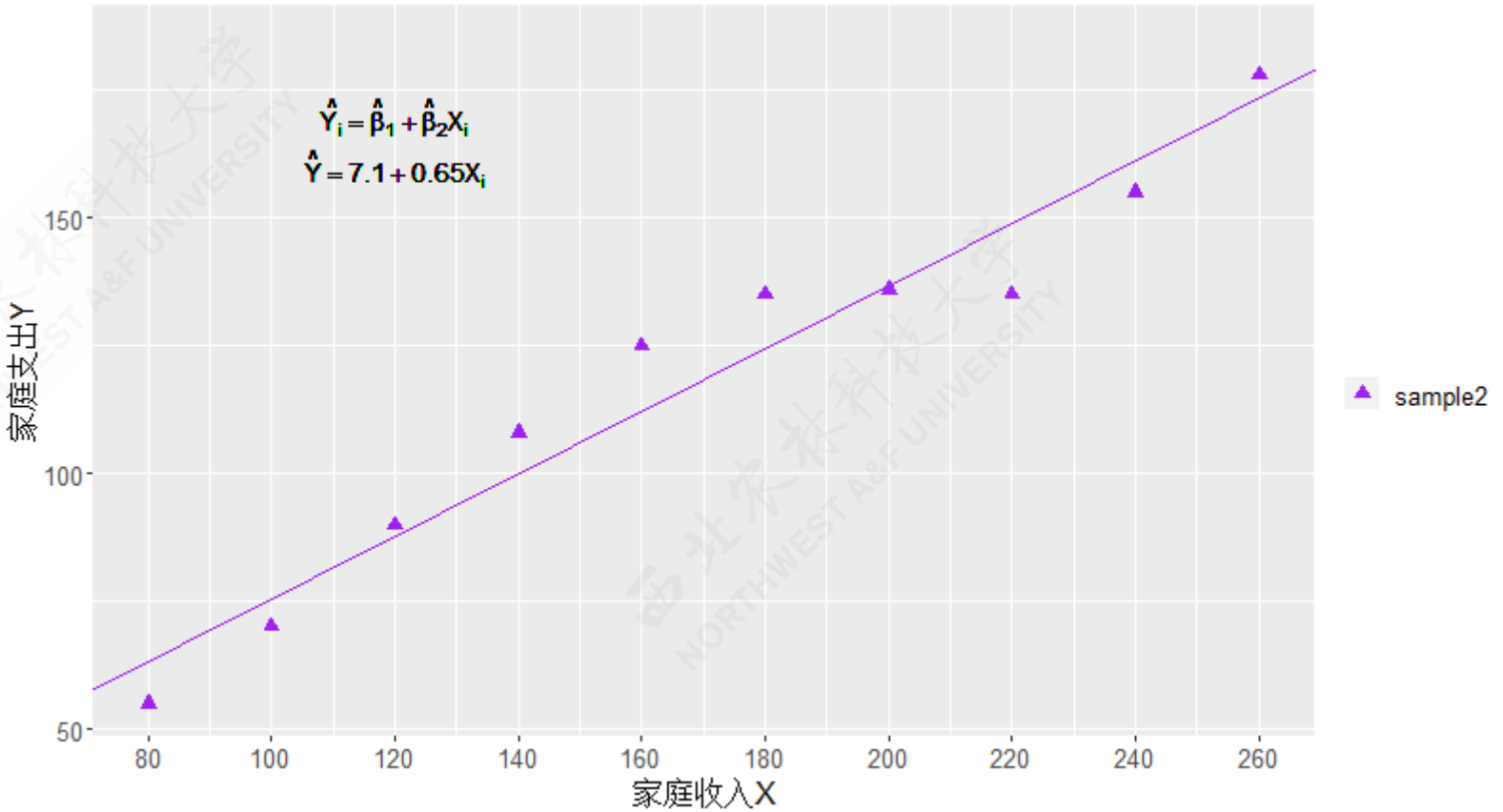
var	n1	n2	n3	n4	n5	n6	n7	n8	n9	n10
X	80	100	120	140	160	180	200	220	240	260
Y	55	70	90	108	125	135	136	135	155	178

（示例）第二份随机样本：数据



var	n1	n2	n3	n4	n5	n6	n7	n8	n9	n10
X	80	100	120	140	160	180	200	220	240	260
Y	55	70	90	108	125	135	136	135	155	178

（示例）第二份随机样本：SRL



var	n1	n2	n3	n4	n5	n6	n7	n8	n9	n10
X	80	100	120	140	160	180	200	220	240	260
Y	55	70	90	108	125	135	136	135	155	178

（示例）第二份随机样本：SRS

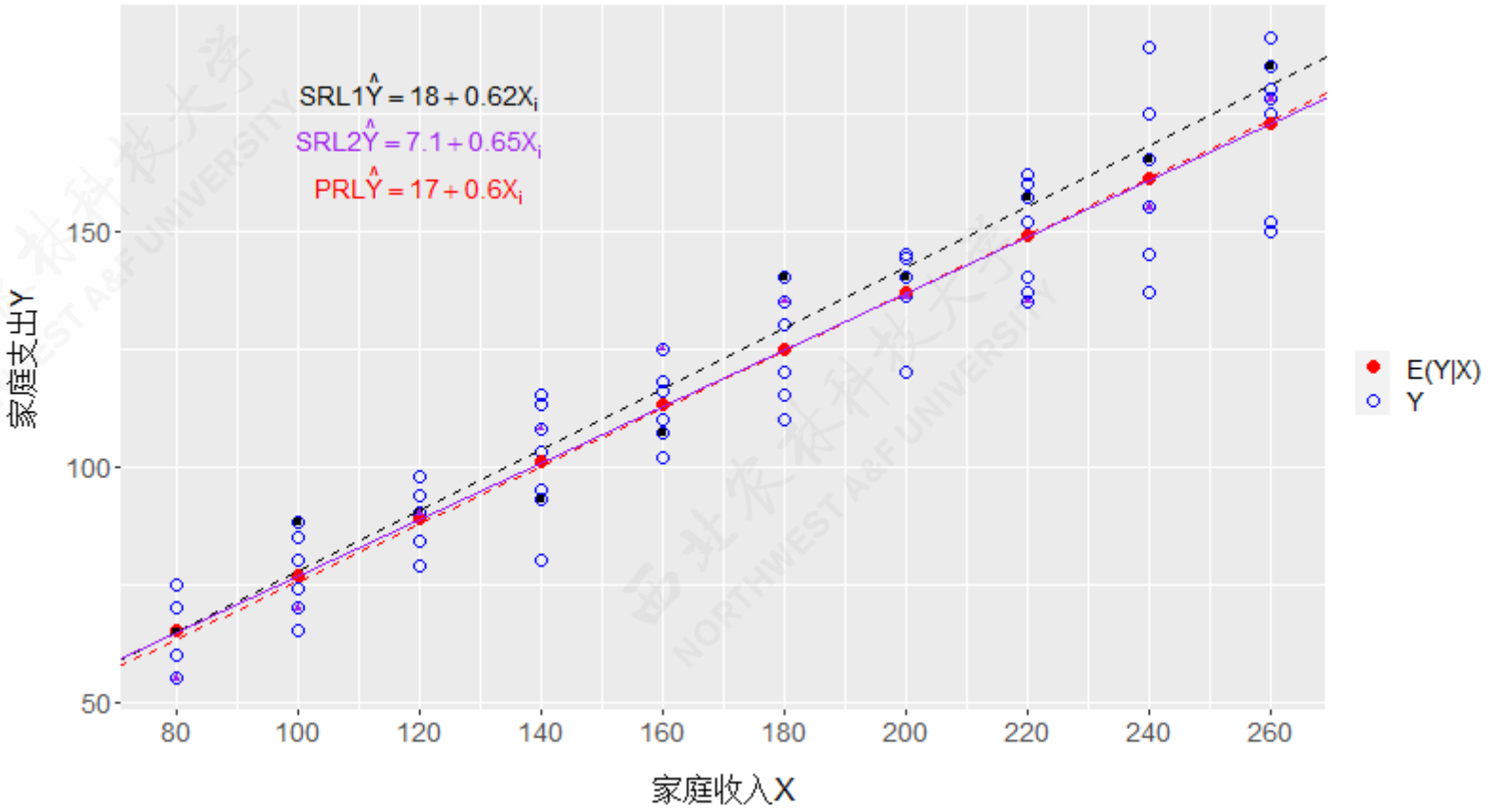
根据第二份随机样本拟合得到的**样本回归函数**SRF：

$$\hat{Y} = +14.59 + 0.61X$$

样本数据如下：

var	n1	n2	n3	n4	n5	n6	n7	n8	n9	n10
X	80	100	120	140	160	180	200	220	240	260
Y	55	70	90	108	125	135	136	135	155	178

（示例）两份样本同时出现：


SRL1 $\hat{Y} = 18 + 0.62X_i$
SRL2 $\hat{Y} = 7.1 + 0.65X_i$
PRL $\hat{Y} = 17 + 0.6X_i$
家庭支出Y
150
100
50
80
100
120
140
160
180
200
220
240
260
家庭收入X
E(Y|X)
Y

重要概念：样本回归模型（SRM）

样本回归模型（Sample Regression Model，SRM）：把样本回归函数表现为**“随机”**形式。

- 如果样本回归函数为隐函数，则样本回归模型可记为：

$$Y_i = g(X_i) + e_i$$

- 如果样本回归函数表现为直线，则样本回归模型可记为：

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i \qquad \text{(SRM\_L)}$$

其中，e_i表示残差（Residual）

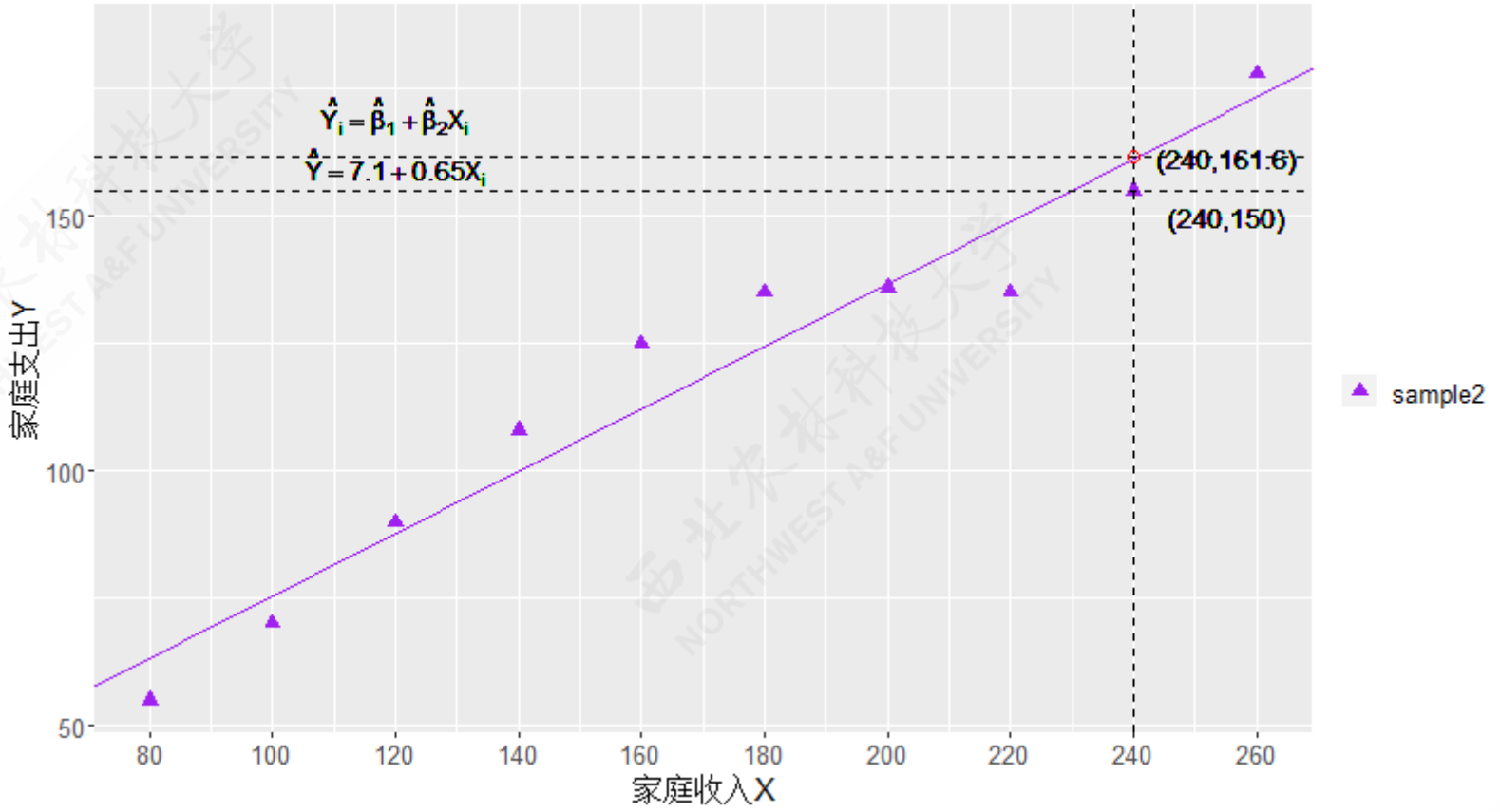
重要概念：残差

残差（Residual）：

- 定义：是样本回归函数与Y的样本观测值之间的离差。
- 记号：

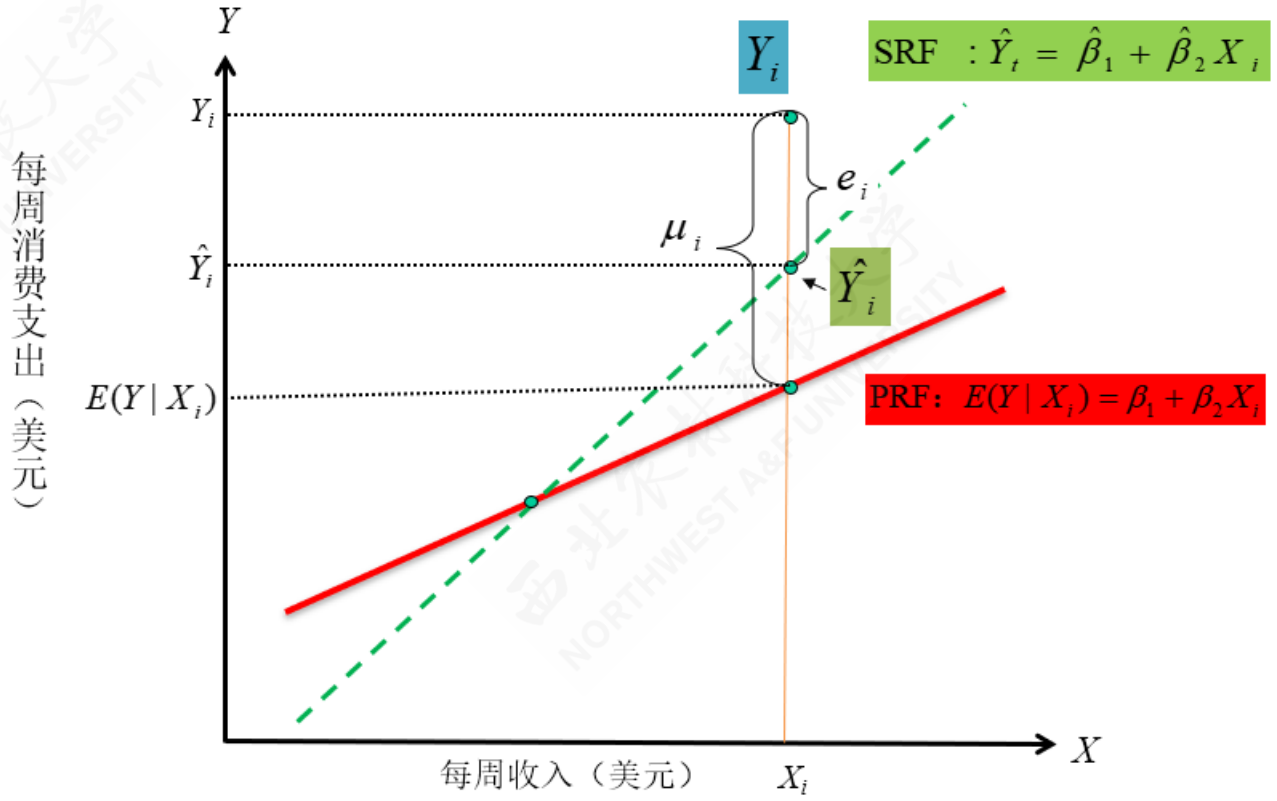
$$
\begin{aligned}
e_i &= Y_i - \hat{Y}_i \\
&= Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_i)
\end{aligned}
$$

重要概念：理解SRF和SRM的关系



给定 $x_i = 240$，样本2的观测值 $Y_i = 240$；拟合值 $\hat{Y}_i = 161.6$；残差 $e_i = Y_i - \hat{Y}_i =$ -6.6。

重要概念：样本回归与总体回归的比较



为何不同？继承性和变异性

重要概念：样本回归与总体回归的比较

总体回归函数PRF:

$$E(Y|X_i) = \beta_1 + \beta_2 X_i \qquad \text{(PRF)}$$

总体回归模型PRM:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad \text{(PRM)}$$

样本回归函数SRF:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i \qquad \text{(SRF)}$$

样本回归模型SRM:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i \qquad \text{(SRM)}$$

思考：

- PRF无法直接观测，只能用SRF近似替代
- 估计值与观测值之间存在偏差
- SRF又是怎样决定的呢?

重要概念：样本回归与总体回归的比较

总结：

- 随机抽样数据继承了总体的特征。
- 利用随机样本进行数据拟合是对总体规律的“反向追踪”。
- 样本回归模型中的残差是拟合不完全的产物。

思考：

- 怎样来判定对随机样本的一次数据拟合是更优的？
- 存不存在一种“最优”的拟合方法？

课后作业：

- 请把162名同学的拟合线进行平均化处理（截距和斜率取均值），绘制得到一条“回归线”。
- 你认为是这根平均化的“回归线”与真相更逼近么？

5.3 OLS方法与参数估计

普通最小二乘法（OLS）

参数估计

估计精度

区间估计

普通最小二乘法（OLS）：引子

我们如何估计回归函数中的系数？

总体回归：

$$\begin{cases} E(Y|X_i) = \beta_1 + \beta_2 X_i & \text{(PRF)} \\ Y_i = \beta_1 + \beta_2 X_i + u_i & \text{(PRM)} \end{cases}$$

样本回归：

$$\begin{cases} \hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i & \text{(SRF)} \\ Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i & \text{(SRM)} \end{cases}$$

首先需要回答的问题是，我们该如何估计得出样本回归函数中的系数？事实上，方法有多种多样：

- 图解法：比较粗糙，但提供了基本的视觉认知
- 最小二乘法(order lease squares, OLS)：最常用的方法
- 最大似然法(maximum likelihood, ML)
- 矩估计方法(Moment method, MM)

普通最小二乘法（OLS）：回顾和比较

总体回归函数PRF:

$$E(Y|X_i) = \beta_1 + \beta_2 X_i$$

总体回归模型PRM:

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

样本回归函数SRF:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

样本回归模型SRM:

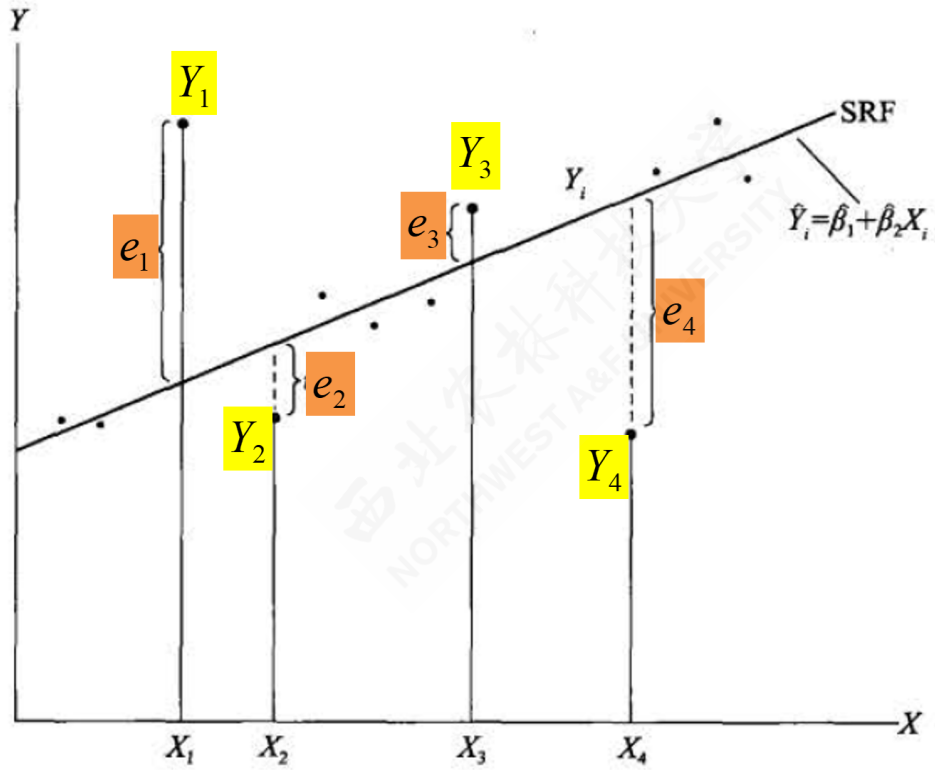
$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$$

思考：

- PRF无法直接观测，只能用SRF近似替代
- 估计值与观测值之间存在偏差
- SRF又是怎样决定的呢?

普通最小二乘法（OLS）：原理

认识普通最小二乘法的原理：一个图示



最小二乘法的原理

普通最小二乘法（OLS）：原理

OLS的基本原理：残差平方和最小化。

$$
\begin{aligned}
e_i &= Y_i - \hat{Y}_i \\
&= Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_i)
\end{aligned}
$$

$$
\begin{aligned}
Q &= \sum e_i^2 \\
&= \sum (Y_i - \hat{Y}_i)^2 \\
&= \sum \left(Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_i)\right)^2 \\
&\equiv f(\hat{\beta}_1, \hat{\beta}_2)
\end{aligned}
$$

$$
Min(Q) = Min\left(f(\hat{\beta}_1, \hat{\beta}_2)\right)
$$

（示例）普通最小二乘法（OLS）的一个数值试验

假设存在下面所示的4组观测值 (X_i, Y_i)：

X_i	Y_i
(1)	(2)
1	4
4	5
5	7
6	12
Sum:	

数值试验：数据

（示例）普通最小二乘法（OLS）的一个数值试验

假设随便猜想了如下两个SRF，完成下表计算，并分析哪个SRF给出的 $(\hat{\beta}_1, \hat{\beta}_2)$ 要更好？

$$SRF1:\ \hat{Y}_{1i} = \hat{\beta}_1 + \hat{\beta}_2 X_i = 1.572 + 1.357 X_i$$
$$SRF2:\ \hat{Y}_{2i} = \hat{\beta}_1 + \hat{\beta}_2 X_i = 3.0 + 1.0 X_i$$

X_i	Y_i	$\hat{Y}_{1i}$	e_{1i}	e_{1i}^2	$\hat{Y}_{2i}$	e_{2i}	e_{2i}^2
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	4	2.929	1.071	1.147	4	0	0
4	5	7.000	2.000	4.000	7	-2	4
5	7	8.357	1.357	1.841	8	-1	1
6	12	9.714	2.285	5.226	9	3	9
Sum:			0.000	12.214		0	14

参数估计：回归参数的OLS点估计

最小化求解：

$$
\begin{aligned}
Min(Q) &= Min\left(f(\hat{\beta}_1,\hat{\beta}_2)\right) \\
&= Min\left(\sum\left(Y_i-(\hat{\beta}_1+\hat{\beta}_2X_i)\right)^2\right) \\
&= Min\sum\left(Y_i-\hat{\beta}_1-\hat{\beta}_2X_i\right)^2
\end{aligned}
$$

方程组变形，得到**正规方程组**：

$$
\begin{cases}
\sum\left[\hat{\beta}_1-(Y_i-\hat{\beta}_2X_i)\right]=0 \\
\sum\left[X_i^2\hat{\beta}_2-(Y_i-\hat{\beta}_1)X_i\right]=0
\end{cases}
$$

$$
\begin{cases}
\sum Y_i-n\hat{\beta}_1-(\sum X_i)\hat{\beta}_2=0 \\
\sum X_iY_i-(\sum X_i)\hat{\beta}_1-(\sum X_i^2)\hat{\beta}_2=0
\end{cases}
$$

参数估计：回归参数的OLS点估计

进而得到回归系数的计算公式1（Favorite Five，FF）：

$$\begin{cases} \hat{\beta}_2 = \dfrac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2} \\ \hat{\beta}_1 = \dfrac{n\sum X_i^2 Y_i - \sum X_i \sum X_i Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2} \end{cases} \quad \text{(FF solution)}$$

参数估计：回归参数的OLS点估计

此外我们也可以得到如下的离差公式(favorite five，ff)

$$\begin{cases} \hat{\beta}_2 = \dfrac{\sum x_i y_i}{\sum x_i^2} \\ \hat{\beta}_1 = \bar{Y}_i - \hat{\beta}_2 \bar{X}_i \end{cases} \quad \text{(ff solution)}$$

其中离差计算 $x_i = X_i - \bar{X};\ y_i = Y_i - \bar{Y}$。

以下式子为什么是等价的？你能推导出来么？

$$\begin{cases} \sum x_i y_i = \sum \left[(X_i - \bar{X})(Y_i - \bar{Y}) \right] = \sum X_i Y_i - \frac{1}{n} \sum X_i \sum Y_i \\ \sum x_i^2 = \sum (X_i - \bar{X})^2 \qquad\qquad = \sum X_i^2 - \frac{1}{n} \left(\sum X_i \right)^2 \end{cases}$$

参数估计：随机干扰项参数的OLS点估计

PRM公式变形：

$$\left.\begin{aligned} Y_i &= \beta_1 - \beta_2 X_i + u_i \ (\text{PRM}) \Rightarrow \\ \hat{Y} &= \beta_1 - \beta_2 \bar{X} + \bar{u} \end{aligned}\right\} \Rightarrow$$
$$y_i = \beta_2 x_i + (u_i - \bar{u})$$

残差公式变形：

$$\left.\begin{aligned} e_i &= y_i - \hat{\beta}_2 x_i \\ e_i &= \beta_2 x_i + (u_i - \bar{u}) - \hat{\beta}_2 x_i \end{aligned}\right\} \Rightarrow$$
$$e_i = -(\hat{\beta}_2 - \beta_2) x_i + (u_i - \hat{u})$$

参数估计：随机干扰项参数的OLS点估计

求解残差平方和：

$$\sum e_i^2 = (\hat{\beta}_2 - \beta_2)^2 \sum x_i^2 + \sum (u - \bar{u})^2 - 2(\hat{\beta}_2 - \beta_2) \sum x_i (u - \bar{u})$$

求残差平方和的期望：

$$\begin{aligned} E(\sum e_i^2) &= \sum x_i^2 E\left[(\hat{\beta}_2 - \beta_2)^2\right] + E\left[\sum (u - \bar{u})^2\right] \\ &\quad + 2E\left[(\hat{\beta}_2 - \beta_2) \sum x_i (u - \bar{u})\right] \\ &\equiv A + B + C \\ &= \sigma^2 + (n-1)\sigma^2 - 2\sigma^2 \\ &= (n-2)\sigma^2 \end{aligned}$$

参数估计：随机干扰项参数的OLS点估计

回归误差方差（Deviation of Regression Error）：

- 采用OLS方法下，总体回归模型PRM中随机干扰项 u_i的总体方差的无偏估计量，记为 $E(\sigma^2) \equiv \hat{\sigma}^2$，简单地记为 $\hat{\sigma}^2$。

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$$

回归误差标准差（Standard Deviation of Regression Error）：有时候也记为**se**。

$$\hat{\sigma} = \sqrt{\frac{\sum e_i^2}{n-2}}$$

（附录）A过程证明

$$
\begin{aligned}
A &= \sum x_i^2 E\left[(\hat{\beta}_2 - \beta_2)^2\right] \\
&= \sum \left[x_i^2 \cdot var(\hat{\beta}_2)\right] \\
&= var(\hat{\beta}_2) \cdot \sum x_i^2 \\
&= \frac{\sigma^2}{\sum x_i^2} \cdot \sum x_i^2 \\
&= \sigma^2
\end{aligned}
$$

（附录）B过程证明

$$
\begin{aligned}
B = E\left[\sum (u-\bar{u})^2\right] &= E(\sum u_i^2) - 2E\left[\sum (u_i\bar{u})\right] + nE(\bar{u}^2) \\
&= n\cdot Var(u_i) - 2E\left[\sum (u_i \cdot \frac{\sum u_i}{n})\right] + nE(\frac{\sum u_i}{n})^2 \\
&= n\sigma^2 - 2E\left[\frac{\sum u_i}{n}\sum u_i\right] + E\left[\frac{(\sum u_i)^2}{n}\right] \\
&= n\sigma^2 - E\left[(\sum u_i)^2/n\right] = n\sigma^2 - \frac{E(u_i^2)+E(u_2^2)+\cdots+E(u_n^2))}{n} \\
&= n\sigma^2 - \frac{nVaru_i}{n} = n\sigma^2 - \sigma^2 = (n-1)\sigma^2
\end{aligned}
$$

（附录）C过程证明

$$
\begin{aligned}
C &= -2E\left[(\hat{\beta}_2-\beta_2)\sum x_i(u_i-\bar{u})\right] \\
&= -2E\left[\frac{\sum x_iu_i}{\sum x_i^2}\left(\sum x_iu_i-\bar{u}\sum x_i\right)\right] \\
&= -2E\left[\frac{\left(\sum x_iu_i\right)^2}{\sum x_i^2}\right] \\
&= -2E\left[(\hat{\beta}_2-\beta_2)^2\right]=-2\sigma^2
\end{aligned}
$$

- 其中：

$$
\hat{\beta}_2=\sum k_iY_i=\sum k_i(\beta_1+\beta_2X_i+u_i)=\beta_1\sum k_i+\beta_2\sum k_iX_i+\sum k_iu_i=\beta_2+\sum k_iu_i
$$

$$
\hat{\beta}_2-\beta_2=\sum k_iu_i=\frac{\sum x_iu_i}{\sum x_i^2}
$$

（案例）计算表的和的

obs	X_i	Y_i	X_iY_i	X_i^2	Y_i^2	x_i	y_i	x_iy_i	x_i^2	y_i^2
1	6.00	4.46	26.74	36.00	19.86	-6.00	-4.22	25.31	36.00	17.79
2	7.00	5.77	40.39	49.00	33.29	-5.00	-2.90	14.52	25.00	8.44
3	8.00	5.98	47.83	64.00	35.74	-4.00	-2.70	10.78	16.00	7.27
4	9.00	7.33	65.99	81.00	53.75	-3.00	-1.34	4.03	9.00	1.80
5	10.00	7.32	73.18	100.00	53.56	-2.00	-1.36	2.71	4.00	1.84
6	11.00	6.58	72.43	121.00	43.35	-1.00	-2.09	2.09	1.00	4.37
7	12.00	7.82	93.82	144.00	61.12	0.00	-0.86	-0.00	0.00	0.73
8	13.00	7.84	101.86	169.00	61.39	1.00	-0.84	-0.84	1.00	0.70
9	14.00	11.02	154.31	196.00	121.49	2.00	2.35	4.70	4.00	5.51
10	15.00	10.67	160.11	225.00	113.93	3.00	2.00	6.00	9.00	4.00
11	16.00	10.84	173.38	256.00	117.42	4.00	2.16	8.65	16.00	4.67
12	17.00	13.62	231.46	289.00	185.37	5.00	4.94	24.70	25.00	24.41
13	18.00	13.53	243.56	324.00	183.09	6.00	4.86	29.14	36.00	23.58
sum	156.00	112.77	1485.04	2054.00	1083.38	0.00	0.00	131.79	182.00	105.12

公式1:（Favorite Five，FF形式）

$$\hat{\beta}_2 = \frac{n\sum X_iY_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2}$$

$$= \frac{13 * 1485.04 - 156 * 112.771}{13 * 2054 - 156^2} = 0.7241$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X} = 8.6747 - 0.7241 * 12 = -0.0145$$

公式2：（离差形式，favorite five，ff形式）

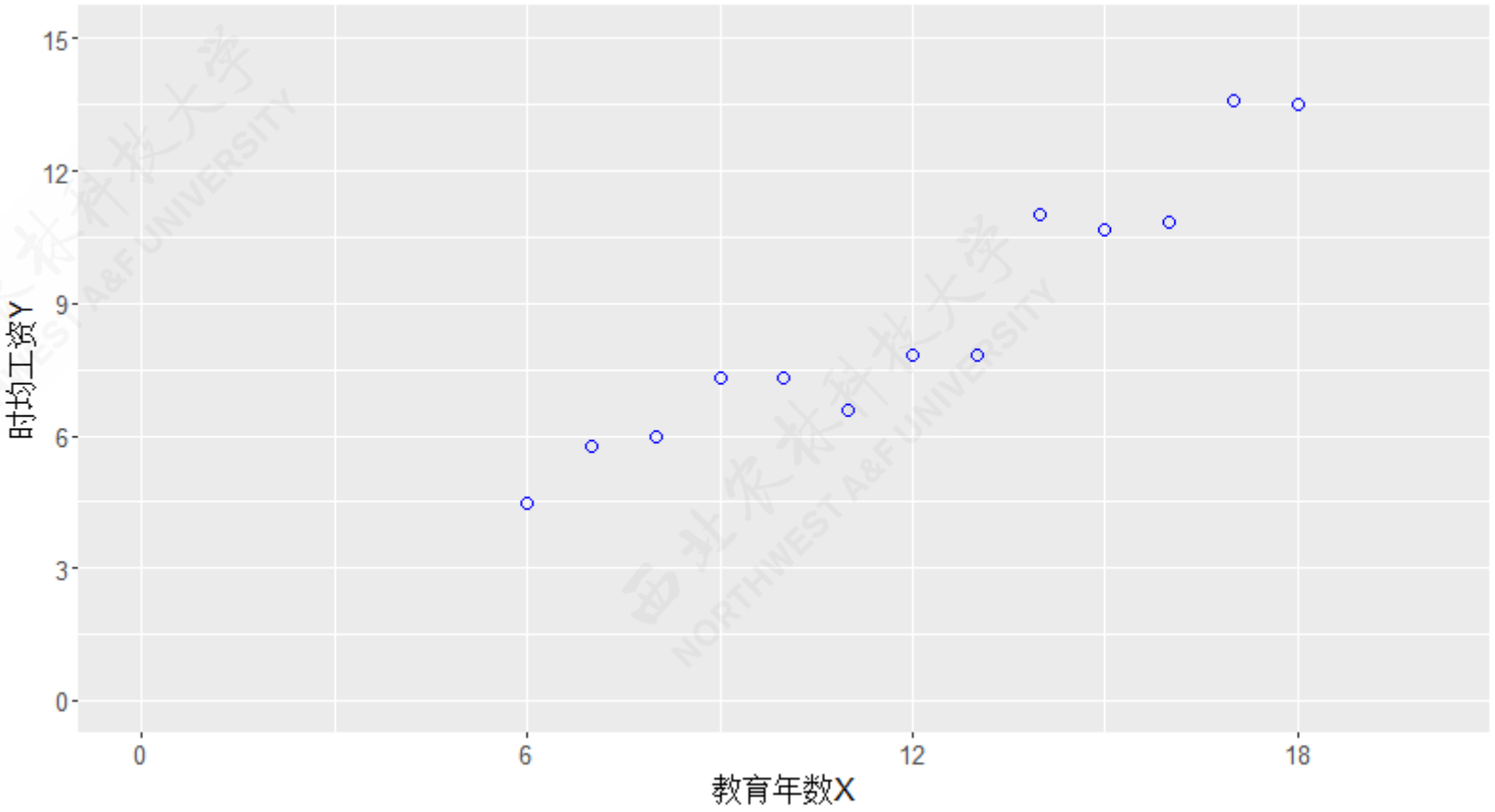
$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{131.786}{182} = 0.7241$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = 8.6747 - 0.7241 * 12 = -0.0145$$

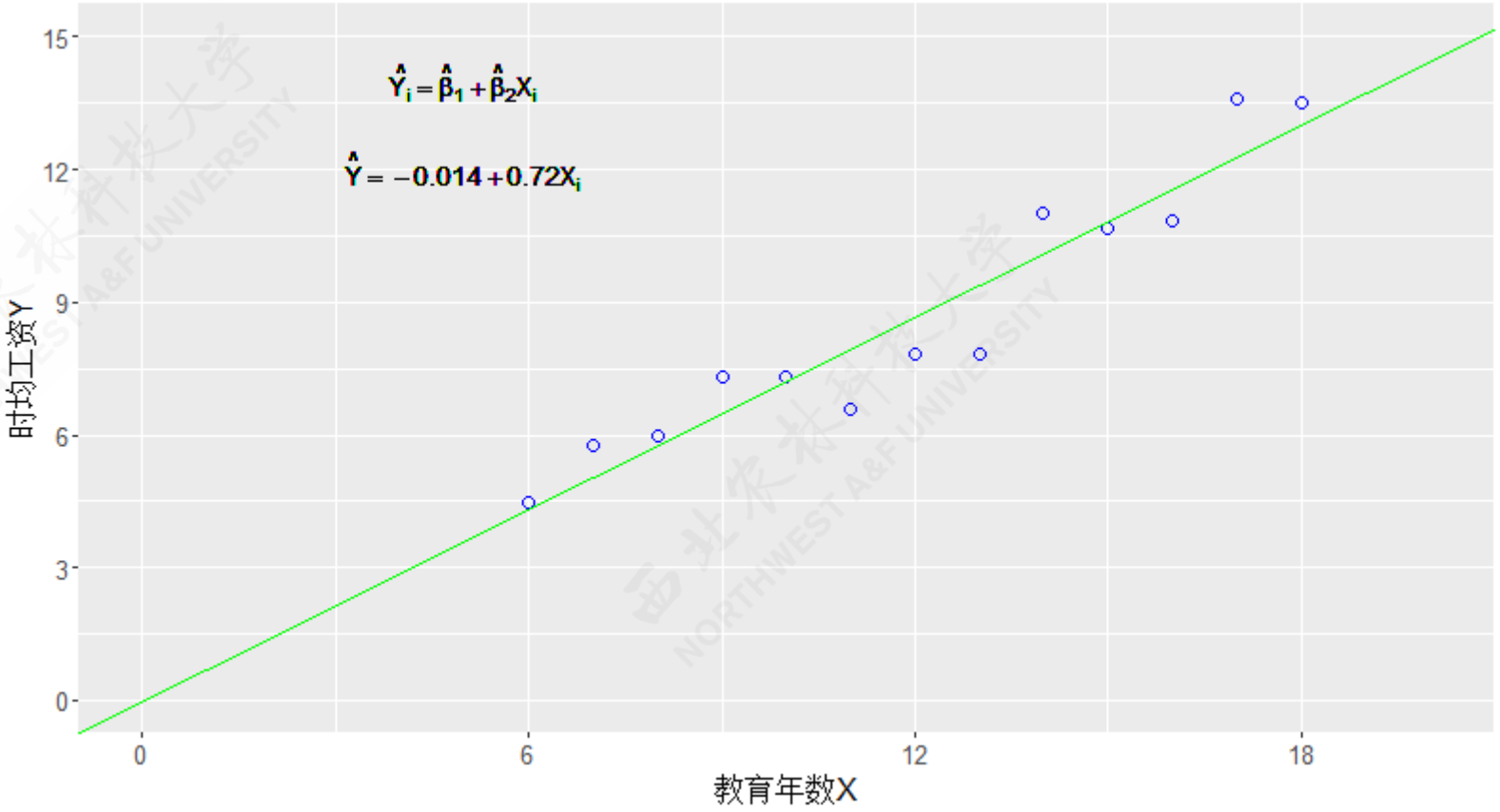
（案例）样本回归方程SRF

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i = -0.0145 + 0.7241 X_i$$

（案例）样本回归线SRL



（案例）样本回归线SRL


$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$
$\hat{Y} = -0.014 + 0.72 X_i$
时均工资Y
教育年数X
0
3
6
9
12
15
0
6
12
18

（案例）计算得到拟合值和残差

obs	`$X_i$`	`$Y_i$`	`$\hat{Y}_i$`	`$e_i$`
1	6	4.5	4.3	0.127
2	7	5.8	5.1	0.716
3	8	6.0	5.8	0.200
4	9	7.3	6.5	0.829
5	10	7.3	7.2	0.092
6	11	6.6	8.0	-1.366
7	12	7.8	8.7	-0.857
8	13	7.8	9.4	-1.564
9	14	11.0	10.1	0.899
10	15	10.7	10.8	-0.173
11	16	10.8	11.6	-0.735
12	17	13.6	12.3	1.320
13	18	13.5	13.0	0.512
sum	156	112.8	112.8	0.000

根据以上样本回归方程，可以计算得到 Y_i的回归拟合值 $\hat{Y}_i$，以及回归残差 e_i。

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$
$$e_i = Y_i - \hat{Y}_i$$

（案例）计算回归误差方差和标准差

回归误差方差 $\hat{\sigma}^2$

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{(n-2)} = \frac{9.693}{11} = 0.8812$$

回归误差标准差 $\hat{\sigma}$:

$$\hat{\sigma} = \sqrt{\frac{\sum e_i^2}{(n-2)}} = \sqrt{0.8812} = 0.9387$$

OLS参数估计：“估计值”与“估计量”

理解OLS方法下的“估计值”与“估计量”

回归系数的计算公式1（Favorite Five，FF）：

$$\begin{cases} \hat{\beta}_2 = \dfrac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2} \\ \hat{\beta}_1 = \dfrac{n\sum X_i^2 Y_i - \sum X_i \sum X_i Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2} \end{cases} \quad \text{(FF solution)}$$

- 如果给出的参数估计结果是由一个具体样本资料计算出来的，它是一个“估计值”，或者“点估计”，是参数估计量的一个具体数值；
- 如果把上式看成参数估计的一个表达式，那么，则它是(X_i, Y_i)的函数，而Y_i是随机变量，所以参数估计也是随机变量，在这个角度上，称之为“估计量”。

OLS参数估计：SRF和SRM的特征

OLS估计量是纯粹由可观测的(即样本)量(指X和Y)表达的，因此它们很容易计算。

它们是点估计量(point estimators)，即对于给定样本，每个估计量仅提供有关总体参数的一个(点)值*。

一旦从样本数据得到OLS估计值，便容易画出样本回归线。

注：我们以后还将考虑区间估计量(interval Estimators)

OLS参数估计：SRF和SRM的特征

- 特征1：样本回归线一定会经过样本均值点 $(\bar{X}, \bar{Y})$：

$$\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}$$

- 特征2： Y_i的**估计值**($\hat{Y}_i$)的均值($\bar{\hat{Y}_i}$)等于Y的样本均值($\bar{Y}$)

$$\begin{aligned} \hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 \bar{X} \\ &= (\bar{Y} - \hat{\beta}_2 \bar{X}) + \hat{\beta}_2 X_i \\ &= \bar{Y} - \hat{\beta}_2 (X_i - \bar{X}) \end{aligned}$$

$$\Rightarrow 1/n \sum \hat{Y}_i = 1/n \sum \bar{Y} - \hat{\beta}_2 (X_i - \bar{X})$$

$$\Rightarrow \bar{\hat{Y}}_i = \bar{Y}$$

OLS参数估计：SRF和SRM的特征

- 特征3：残差的均值($\bar{e}_i$)为零：

$$\sum \left[\hat{\beta}_1 - (Y_i - \hat{\beta}_2 X_i)\right] = 0$$
$$\sum \left[Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)\right] = 0$$
$$\sum (Y_i - \hat{Y}_i) = 0$$
$$\sum e_i = 0$$
$$\bar{e}_i = 0$$

OLS参数估计：SRF和SRM的特征

- 特征4：SRM和SRF可以写成离差形式：

$$\left.\begin{aligned} Y_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i \\ \bar{Y} &= \hat{\beta}_1 + \hat{\beta}_2 \bar{X} \end{aligned}\right\} \Rightarrow$$

$$Y_i - \bar{Y} = \hat{\beta}_2 (X_i - \bar{X}) + e_i \Rightarrow$$

$$y_i = \hat{\beta}_2 x_i + e_i \qquad \text{(SRM-dev)}$$

$$\left.\begin{aligned} \hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i \\ \bar{Y} &= \hat{\beta}_1 + \hat{\beta}_2 \bar{X} \end{aligned}\right\} \Rightarrow$$

$$\hat{Y}_i - \bar{Y} = \hat{\beta}_2 (X_i - \bar{X}) \Rightarrow$$

$$\hat{y}_i = \hat{\beta}_2 x_i \qquad \text{(SRF-dev)}$$

OLS参数估计：SRF和SRM的特征

- 特征5：残差(e_i)和 Y_i的拟合值($\hat{Y}_i$)不相关

$$
\begin{aligned}
Cov(e_i, \hat{Y}_i) &= E\left[(e_i - E(e_i)) \cdot \left(\hat{Y}_i - E(\hat{Y}_i)\right)\right] = E(e_i \cdot \hat{y}_i) \\
&= \sum (e_i \cdot \hat{\beta}_2 x_i) \\
&= \sum \left[(y_i - \hat{\beta}_2 x_i) \cdot \hat{\beta}_2 x_i\right] \\
&= \hat{\beta}_2 \sum \left[(y_i - \hat{\beta}_2 x_i) \cdot x_i\right] \\
&= \hat{\beta}_2 \sum \left[(y_i x_i - \hat{\beta}_2 x_i^2)\right] \\
&= \hat{\beta}_2 \sum x_i y_i - \hat{\beta}_2^2 \sum x_i^2 \qquad \Leftarrow \hat{\beta}_2 = \frac{\sum x_i y_i}{x_i^2} \\
&= \hat{\beta}_2^2 \sum x_i^2 - \hat{\beta}_2{}^2 \sum x_i^2 = 0
\end{aligned}
$$

- 特征6：残差(e_i)和自变量(X_i)不相关

OLS参数估计：离差公式

- 离差定义与符号：

$$
\begin{aligned}
x_i &= X_i - \bar{X} \\
y_i &= Y_i - \bar{Y} \\
\hat{y}_i &= \hat{Y}_i - \bar{\hat{Y}}_i = \hat{Y}_i - \bar{Y}
\end{aligned}
$$

- PRM及其离差形式：

$$
\left.
\begin{aligned}
Y_i &= \beta_1 + \beta_2 X_i + u_i \\
\bar{Y} &= \beta_1 + \beta_2 \bar{X} + \bar{u}
\end{aligned}
\right\} \Rightarrow
$$

$$
Y_i - \bar{Y} = \beta_2 x_i + (u_i - \bar{u}) \Rightarrow
$$

$$
y_i = \hat{\beta}_2 x_i + (u_i - \bar{u}) \qquad \text{(PRM-dev)}
$$

OLS参数估计：离差公式

- SRM及其离差形式：

$$\left.\begin{aligned} Y_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i \\ \bar{Y} &= \hat{\beta}_1 + \hat{\beta}_2 \bar{X} \end{aligned}\right\} \Rightarrow$$

$$Y_i - \bar{Y} = \hat{\beta}_2 (X_i - \bar{X}) + e_i \Rightarrow$$

$$y_i = \hat{\beta}_2 x_i + e_i$$

- SRF及其离差形式：

$$\left.\begin{aligned} \hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i \\ \bar{Y} &= \hat{\beta}_1 + \hat{\beta}_2 \bar{X} \end{aligned}\right\} \Rightarrow$$

$$\hat{Y}_i - \bar{Y} = \hat{\beta}_2 (X_i - \bar{X}) \Rightarrow$$

$$\hat{y}_i = \hat{\beta}_2 x_i$$

- 残差的离差形式：

$$y_i = \hat{\beta}_2 x_i + e_i \qquad \text{(SRM-dev)} \Rightarrow$$

$$e_i = y_i - \hat{\beta}_2 x_i \qquad \text{(residual-dev)}$$

OLS参数估计：思考与讨论

内容小结：

- 普通最小二乘方法（OLS）采用“铅垂线距离平方和最小化”的思想，来拟合一条样本回归线，进而求解出模型参数估计量。
- 大家需要很熟练地记住OLS参数估计量公式，以及它们的几大重要特征！

思考讨论：

- OLS采用的“铅垂线距离平方和最小化”这一方案，凭什么它被奉为计量分析的经典方法？你觉得还有其他可行替代方案么？
- 回归标准误差 se的现实含义是什么？回归参数估计与随机干扰项的方差估计有什么内在联系么？
- OLS方法的几个特征，是不是使它“天生丽质”、“娘胎里生下来就含着金钥匙”？为什么能这么说？

估计精度：引子

我们已经使用OLS方法分别得到总体回归模型(PRM)的3个重要参数（实际不止3个）的点估计量：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2}; \quad \hat{\beta}_1 = \bar{Y}_i - \hat{\beta}_2 \bar{X}_i; \quad \hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$$

> 问题是：我们如何知道OLS方法点估计量是否可靠？OLS方法的点估计量是否稳定？OLS方法的点估计量是否可信？

因此，我们需要找到一种表达OLS方法估计稳定性或估计精度的指标！

- 点估计量的**方差**（variance）和**标准差**（standard deviation）就是衡量估计稳定性或估计精度的一类重要指标！

估计精度：斜率系数的方差和样本方差

斜率系数（$\hat{\beta}_2$）的**总体方差**（$\sigma^2_{\hat{\beta}_2}$）和**总体标准差**（$\sigma_{\hat{\beta}_2}$）：

$$Var(\hat{\beta}_2) \equiv \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2}{\sum x_i^2}$$

$$\sigma_{\hat{\beta}_2} = \sqrt{\frac{\sigma^2}{\sum x_i^2}}$$

- 其中，$Var(u_i) \equiv \sigma^2$表示随机干扰项u_i的总体方差。

斜率系数（$\hat{\beta}_2$）的**样本方差**（$S^2_{\hat{\beta}_2}$）和**样本标准差**（$S_{\hat{\beta}_2}$）：

$$S^2_{\hat{\beta}_2} = \frac{\hat{\sigma}^2}{\sum x_i^2}$$

$$S_{\hat{\beta}_2} = \sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2}}$$

- 其中，$E(\sigma^2) = \hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$表示对随机干扰项（$u_i$）的总体方差的**无偏估计量**。

（附录）证明过程I

步骤1 $\hat{\beta}_2$的变形：

$$
\begin{aligned}
\hat{\beta}_2 &= \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum \left[x_i (Y_i - \bar{Y})\right]}{\sum x_i^2} \\
&= \frac{\sum x_i Y_i - \sum x_i \bar{Y}}{\sum x_i^2} \\
&= \frac{\sum x_i Y_i - \bar{Y} \sum x_i}{\sum x_i^2} && \leftarrow \left[\sum x_i = \sum (X_i - \bar{X}) = 0\right] \\
&= \sum \left(\frac{x_i}{\sum x_i^2} \cdot Y_i \right) && \leftarrow \left[k_i \equiv \frac{x_i}{\sum x_i^2}\right] \\
&= \sum k_i Y_i
\end{aligned}
$$

- 其中，$k_i \equiv \frac{x_i}{\sum x_i^2}$。

（附录）证明过程2

步骤2： 计算 $\hat{\beta}_2$ 的**总体方差**（$\sigma^2_{\hat{\beta}_2}$）：

$$
\begin{aligned}
\sigma^2_{\hat{\beta}_2} &\equiv Var(\hat{\beta}_2) = Var(\sum k_i Y_i) \\
&= \sum \left(k_i^2 Var(Y_i)\right) \\
&= \sum \left(k_i^2 Var(\beta_1 + \beta_2 X_i + u_i)\right) \\
&= \sum \left(k_i^2 Var(u_i)\right) \qquad \leftarrow \left[k_i \equiv \frac{x_i}{\sum x_i^2}\right] \\
&= \sum \left(\left(\frac{x_i}{\sum x_i^2}\right)^2 \cdot \sigma^2\right) \\
&= \frac{\sigma^2}{\sum x_i^2}
\end{aligned}
$$

> 其中，$Var(u_i) \equiv \sigma^2$ 表示随机干扰项 u_i 的总体方差。

估计精度：截距系数的方差和样本方差

截距系数（$\hat{\beta}_1$）的**总体方差**（$\sigma^2_{\hat{\beta}_1}$）和**总体标准差**（$\sigma_{\hat{\beta}_1}$）：

$$Var(\hat{\beta}_1) \equiv \sigma^2_{\hat{\beta}_1} = \frac{\sum X_i^2}{n} \cdot \frac{\sigma^2}{\sum x_i^2}$$

$$\sigma_{\hat{\beta}_1} = \sqrt{\frac{\sum X_i^2}{n} \cdot \frac{\sigma^2}{\sum x_i^2}}$$

- 其中，$Var(u_i) \equiv \sigma^2$表示随机干扰项u_i的总体方差。

截距系数（$\hat{\beta}_1$）的**样本方差**（$S^2_{\hat{\beta}_1}$）和**样本标准差**（$S_{\hat{\beta}_1}$）：

$$S^2_{\hat{\beta}_1} = \frac{\sum X_i^2}{n} \cdot \frac{\hat{\sigma}^2}{\sum x_i^2}$$

$$S_{\hat{\beta}_1} = \sqrt{\frac{\sum X_i^2}{n} \cdot \frac{\hat{\sigma}^2}{\sum x_i^2}}$$

- 其中，$E(\sigma^2) = \hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$表示对随机干扰项（$u_i$）的总体方差的**无偏估计量**。

步骤1 $\hat{\beta}_1$的变形：

$$
\begin{aligned}
\hat{\beta}_1 &= \bar{Y}_i - \hat{\beta}_2 \bar{X}_i && \leftarrow \left[\hat{\beta}_2 = \sum k_i Y_i\right] \\
&= \frac{1}{n}\sum Y_i - \sum \left(k_i Y_i \cdot \bar{X}\right) \\
&= \sum \left((\frac{1}{n} - k_i \bar{X}) \cdot Y_i \right) && \leftarrow \left[w_i \equiv \frac{1}{n} - k_i \bar{X}\right] \\
&= \sum w_i Y_i
\end{aligned}
$$

- 其中：令 $w_i \equiv \frac{1}{n} - k_i \bar{X}$

（附录）证明过程2

步骤2计算 $\hat{\beta}_1$ 的**总体方差**（$\sigma^2_{\hat{\beta}_1}$）：

$$
\begin{aligned}
\sigma^2_{\hat{\beta}_1} &\equiv Var(\hat{\beta}_1) = Var(\sum w_i Y_i) \\
&= \sum \left(w_i^2 Var(\beta_1 + \beta_2 X_i + u_i) \right) && \leftarrow \left[w_i \equiv \frac{1}{n} - k_i \bar{X} \right] \\
&= \sum \left(\left(\frac{1}{n} - k_i \bar{X} \right)^2 Var(u_i) \right) \\
&= \sigma^2 \cdot \sum \left(\frac{1}{n^2} - \frac{2\bar{X} k_i}{n} + k_i^2 \bar{X}^2 \right) && \leftarrow \left[\sum k_i = \sum \left(\frac{x_i}{\sum x_i^2} \right) = \frac{\sum x_i}{\sum x_i^2} = 0 \right] \\
&= \sigma^2 \cdot \left(\frac{1}{n} + \bar{X}^2 \sum k_i^2 \right) && \leftarrow \left[k_i \equiv \frac{x_i}{\sum x_i^2} \right] \\
&= \sigma^2 \cdot \left(\frac{1}{n} + \bar{X}^2 \sum \left(\frac{x_i}{\sum x_i^2} \right)^2 \right)
\end{aligned}
$$

步骤2计算 $\hat{\beta}_1$ 的**总体方差**（$\sigma^2_{\hat{\beta}_1}$）（续前）：

$$
\begin{aligned}
&= \sigma^2 \cdot \left(\frac{1}{n} + \bar{X}^2 \frac{\sum x_i^2}{\left(\sum x_i^2\right)^2} \right) \\
&= \sigma^2 \cdot \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2} \right) \\
&= \frac{\sum x_i^2 + n\bar{X}^2}{n \sum x_i^2} \cdot \sigma^2 \qquad \leftarrow \left[\sum x_i^2 + n\bar{X}^2 = \sum (X_i - \bar{X})^2 + n\bar{X}^2 = \sum X_i^2 \right] \\
&= \frac{\sum X_i^2}{n} \cdot \frac{\sigma^2}{\sum x_i^2}
\end{aligned}
$$

估计精度：小结与思考

现在做一个**内容小结**：

- 为了衡量OLS方法的点估计量是否稳定或是否可信，我们一般采用方差和标准差指标来表达。
- 大家应熟记**斜率**和**截距**估计量的**总体方差**和**样本方差**最终公式。

请大家**思考**如下问题：

- 总体方差和样本方差都是确定的数么？
- 二者分别受那些因素的影响？二者又有什么联系？
- 证明过程中，约定的 k_i和 w_i，有什么特征？

$$\begin{cases} \sum k_i = 0 \\ \sum k_i X_i = 1 \end{cases} \qquad \begin{cases} \sum w_i = 1 \\ \sum w_i X_i = 0 \end{cases}$$

（案例）计算回归系数的样本方差

对于“教育程度案例”，利用FF-ff计算表，以及我们已算出的如下计算量：

- 回归误差方差：$\hat{\sigma}^2 = 0.8812$。

则可以进一步计算出，回归系数的样本方差的标准差分别为：

$$S^2_{\hat{\beta}_2} = \frac{\hat{\sigma}^2}{\sum x_i^2} = \frac{0.8812}{182} = 0.0048$$

$$S_{\hat{\beta}_2} = \sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2}} = \sqrt{0.0048} = 0.0696$$

$$S^2_{\hat{\beta}_1} = \frac{\sum X_i^2}{n}\frac{\hat{\sigma}^2}{\sum x_i^2} = \frac{2054}{13}\frac{0.8812}{182} = 0.765$$

$$S_{\hat{\beta}_1} = \sqrt{\frac{\sum X_i^2}{n}\frac{\hat{\sigma}^2}{\sum x_i^2}} = \sqrt{0.765} = 0.8746$$

区间估计：斜率系数

$$\hat{\beta}_2 \sim N(\mu_{\hat{\beta}_2}, \sigma^2_{\hat{\beta}_2}) \quad \leftarrow \left[\mu_{\hat{\beta}_2} = \beta_2; \quad \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2}{\sum x_i^2}\right]$$

$$Z = \frac{\left(\hat{\beta}_2 - \beta_2\right)}{\sqrt{\mathrm{var}\left(\hat{\beta}_2\right)}} = \frac{\left(\hat{\beta}_2 - \beta_2\right)}{\sqrt{\sigma^2_{\beta_2}}} = \frac{\hat{\beta}_2 - \beta_2}{\sigma_{\hat{\beta}_2}} = \frac{\left(\hat{\beta}_2 - \beta_2\right)}{\sqrt{\frac{\sigma^2}{\sum x_i^2}}} \quad \leftarrow Z \sim N(0,1)$$

$$T = \frac{\left(\hat{\beta}_2 - \beta_2\right)}{\sqrt{S^2_{\beta_2}}} = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{S^2_{\beta_2}}} = \frac{\hat{\beta}_2 - \beta_2}{S_{\hat{\beta}_2}} \quad \leftarrow T \sim t(n-2)$$

$$S^2_{\hat{\beta}_2} = \frac{\hat{\sigma}^2}{\sum x_i^2}; \quad \hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$$

$$\Pr\left[-t_{\alpha/2,(n-2)} \le \mathrm{T} \le t_{\alpha/2,(n-2)}\right] = 1 - \alpha$$

区间估计：斜率系数

$$\Pr\left[-t_{\alpha/2,(n-2)} \leq \frac{\hat{\beta}_2 - \beta_2}{S_{\hat{\beta}_2}} \leq t_{\alpha/2,(n-2)}\right] = 1-\alpha$$

$$\Pr\left[\hat{\beta}_2 - t_{\alpha/2,(n-2)} \cdot S_{\hat{\beta}_2} \leq \beta_2 \leq \hat{\beta}_2 + t_{\alpha/2,(n-2)} \cdot S_{\hat{\beta}_2}\right] = 1-\alpha$$

因此，β_2的$100(1-\alpha)\%$置信上限和下限分别为：

$$\hat{\beta}_2 \pm t_{\alpha/2} \cdot S_{\hat{\beta}_2}$$

β_2的$100(1-\alpha)\%$置信区间为：

$$\left[\hat{\beta}_2 - t_{\alpha/2} \cdot S_{\hat{\beta}_2}, \quad \hat{\beta}_2 + t_{\alpha/2} \cdot S_{\hat{\beta}_2}\right]$$

区间估计：截距系数

$$\hat{\beta}_1 \sim N(\mu_{\hat{\beta}_1}, \sigma^2_{\hat{\beta}_1}) \quad \leftarrow \left[\mu_{\hat{\beta}_1} = \beta_1; \quad \sigma^2_{\hat{\beta}_1} = \frac{\sum X_i^2}{n}\frac{\sigma^2}{\sum x_i^2}\right]$$

$$Z = \frac{\left(\hat{\beta}_1 - \beta_1\right)}{\sqrt{\mathrm{var}\left(\hat{\beta}_1\right)}} = \frac{\left(\hat{\beta}_1 - \beta_1\right)}{\sqrt{\sigma^2_{\beta_1}}} = \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} = \frac{\left(\hat{\beta}_1 - \beta_1\right)}{\sqrt{\frac{\sum X_i^2}{n} \cdot \frac{\sigma^2}{\sum x_i^2}}} \quad \leftarrow Z \sim N(0,1)$$

$$T = \frac{\left(\hat{\beta}_1 - \beta_1\right)}{S^2_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{S^2_{\beta_1}}} = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} \quad \leftarrow T \sim t(n-2)$$

$$S^2_{\hat{\beta}_1} = \frac{\sum X_i^2}{n} \cdot \frac{\hat{\sigma}^2}{\sum x_i^2}; \quad \hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$$

$$\Pr\left[-t_{\alpha/2,(n-2)} \le \mathrm{T} \le t_{\alpha/2,(n-2)}\right] = 1 - \alpha$$

区间估计：截距系数

$$\Pr\left[-t_{\alpha/2,(n-2)} \leq \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} \leq t_{\alpha/2,(n-2)}\right] = 1-\alpha$$

$$\Pr\left[\hat{\beta}_1 - t_{\alpha/2,(n-2)} \cdot S_{\hat{\beta}_1} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,(n-2)} \cdot S_{\hat{\beta}_1}\right] = 1-\alpha$$

因此，β_1的 $100(1-\alpha)\%$置信上限和下限分别为：

$$\hat{\beta}_1 \pm t_{\alpha/2} \cdot S_{\hat{\beta}_1}$$

β_1的 $100(1-\alpha)\%$置信区间为：

$$\left[\hat{\beta}_1 - t_{\alpha/2} \cdot S_{\hat{\beta}_1}, \quad \hat{\beta}_1 + t_{\alpha/2} \cdot S_{\hat{\beta}_1}\right]$$

区间估计：随机干扰项的方差

$$\chi^2 = (n-2)\frac{\hat{\sigma}^2}{\sigma^2} \quad \leftarrow \quad \chi^2 \sim \chi^2(n-2)$$

$$\Pr\left(\chi^2_{\alpha/2} \le \chi^2 \le \chi^2_{\alpha/2}\right) = 1-\alpha$$

$$\Pr\left(\chi^2_{\alpha/2} \le (n-2)\frac{\hat{\sigma}^2}{\sigma^2} \le \chi^2_{1-\alpha/2}\right) = 1-\alpha$$

$$\Pr\left[(n-2)\frac{\hat{\sigma}^2}{\chi^2_{1-\alpha/2}} \le \sigma^2 \le (n-2)\frac{\hat{\sigma}^2}{\chi^2_{\alpha/2}}\right] = 1-\alpha$$

因此， σ^2的 $100(1-\alpha)\%$为：

$$\left[(n-2)\frac{\hat{\sigma}^2}{\chi^2_{1-\alpha/2}}, \quad (n-2)\frac{\hat{\sigma}^2}{\chi^2_{\alpha/2}}\right]$$

（案例）主模型

我们继续利用样本数据对**教育和工资案例**进行分析。

> **教育和工资案例**的总体回归模型（PRM）如下：

$$
\begin{aligned}
Wage_i &= \beta_1 + \beta_2 Edu_i + u_i \\
Y_i &= \beta_1 + \beta_2 X_i + u_i
\end{aligned}
$$

> **教育和工资案例**的总体回归模型（SRM）如下：

$$
\begin{aligned}
\widehat{Wage}_i &= \hat{\beta}_1 + \hat{\beta}_2 Edu_i + e_i \\
\hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i
\end{aligned}
$$

（案例）相关计算量

我们之前已算出“教育程度案例”中的如下计算量：

- 回归系数：$\hat{\beta}_1 = \text{-}0.0145$；$\hat{\beta}_2 = 0.7241$；$\hat{\sigma}^2 = 0.8812$。
- 回归误差方差：$\hat{\sigma}^2 = 0.8812$。
- 回归系数的样本方差：$S^2_{\hat{\beta}_1} = \frac{\sum X_i^2}{n} \cdot \frac{\hat{\sigma}^2}{\sum x_i^2} = 0.7650$；$S^2_{\hat{\beta}_2} = \frac{\hat{\sigma}^2}{\sum x_i^2} = 0.0048$;
- 回归系数的样本标准差：$S_{\hat{\beta}_1} = 0.8746$；$S_{\hat{\beta}_2} = 0.0696$。

给定 $\alpha = 0.05$，$(1-\alpha)100\% = 95\%$，我们可以查t分布表得到理论参照值：

$t_{\alpha/2}(n-2) = t_{0.05/2}(11) = 2.2010$

（案例）回归系数的区间估计

下面我们进一步计算回归系数的置信区间：

那么，截距参数β_1的95%置信区间为：

$$\begin{aligned} \hat{\beta}_1 - t_{\alpha/2} \cdot S_{\hat{\beta}_1} \quad &\leq \beta_1 \leq \quad \hat{\beta}_1 + t_{\alpha/2} \cdot S_{\hat{\beta}_1} \\ -0.0145 - 2.201 * 0.8746 \quad &\leq \beta_1 \quad \leq -0.0145 + 2.201 * 0.8746 \\ -1.9395 \quad &\leq \beta_1 \quad \leq 1.9106 \end{aligned}$$

那么，斜率参数β_2的95%置信区间为：

$$\begin{aligned} \hat{\beta}_2 - t_{\alpha/2} \cdot S_{\hat{\beta}_2} \quad &\leq \beta_2 \leq \quad \hat{\beta}_2 + t_{\alpha/2} \cdot S_{\hat{\beta}_2} \\ 0.7241 - 2.201 * 0.0696 \quad &\leq \beta_2 \quad \leq 0.7241 + 2.201 * 0.0696 \\ 0.5709 \quad &\leq \beta_2 \quad \leq 0.8772 \end{aligned}$$

（案例）随机干扰项方差的区间估计

- 给定 $\alpha = 0.05,\quad (1-\alpha)100\% = 95\%$
- 查卡方分布表可知：
  - $\chi^2_{\alpha/2}(n-2) = \chi^2_{0.05/2}(11) = \chi^2_{0.025}(11) = 3.8157$
  - $\chi^2_{1-\alpha/2}(n-2) = \chi^2_{1-0.05/2}(11) = \chi^2_{0.975}(11) = 21.9200$

们之前已算出回归误差方差 $\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = 0.8812$ 。因此可以算出 σ^2的95%置信区间为：

$$(n-2)\frac{\hat{\sigma}^2}{\chi^2_{\alpha}} \leq \sigma^2 \leq (n-2)\frac{\hat{\sigma}^2}{\chi^2_{1-\alpha/2}}$$

$$11 * \frac{0.8812}{21.92} \leq \sigma^2 \leq 11 * \frac{0.8812}{3.8157}$$

$$0.4422 \leq \sigma^2 \leq 2.5403$$

5.4 假设检验

两种检验方法

回归系数t检验

方差分解（ANOVA）

模型整体显著性F检验

假设检验：原理和思路

假设检验（Hypothesis Testing）：某一给定的观测或发现与某声称的假设是否相符？进行统计假设检验，就是要制定一套步骤和规则，以使决定接受或拒绝一个虚拟假设（原假设）。

虚拟假设(null hypothesis) —— H_0

- 指定或声称的假设，如 $H_0 : \beta_2 = 0$
- 它是一个等待被挑战的“靶子”！“稻草人”！

备择假设(alternative hypothesis) —— H_1

- 简单的（simple）备择假设，如 $H_1 : \beta_2 = 1.5$
- 复合的（composite）备择假设，如 $H_1 : \beta_2 \neq 1.5$

假设检验的具体方法：

- **置信区间检验**（confidence interval）
- **显著性检验**（test of significance）

假设检验：置信区间检验法（双侧检验）

双侧或双尾检验（Two-sided or Two-Tail Test）

$$H_0 : \beta_2 = 0; \quad H_1 : \beta_2 \neq 0$$

- 假设检验目的：估计的是否与上述相容?
- 决策规则：
  - 构造一个 β_2的 $100(1-\alpha)\%$置信区间。
  - 如果 β_2在 H_0假设下落入此区间，就不拒绝 H_0。
  - 如果它落在此区间之外，就要拒绝 H_0。

对于**斜率参数** β_2的置信区间检验法。

- **步骤1**：给出模型，并提出假设：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

$$H_0 : \beta_2 = 0.5; \quad H_1 : \beta_2 \neq 0.5$$

- **步骤2**：给定 $\alpha = 0.05$，$(1-\alpha)100\% = 95\%$
- **步骤3**：根据前述计算结果，计算斜率参数 β_2的95%置信区间为：

$$\begin{aligned} \hat{\beta}_2 - t_{\alpha/2} \cdot S_{\hat{\beta}_2} \quad &\leq \beta_2 \leq \quad \hat{\beta}_2 + t_{\alpha/2} \cdot S_{\hat{\beta}_2} \\ 0.5709 \quad &\leq \beta_2 \quad \leq 0.8772 \end{aligned}$$

- **步骤4**：那么我们可以对斜率参数 β_2做出如下检验判断：拒绝原假设 H_0，接受 H_1。认为，长期来看很多个区间 [0.5709,0.8772] 有95%的可能性不包含0.5（$\beta_2 \neq 0.5$）。

对于**截距参数** β_1的置信区间检验法。

- **步骤1**：给出模型，并提出假设：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

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

- **步骤2**：给定 $\alpha = 0.05, \quad (1-\alpha)100\% = 95\%$
- **步骤3**：根据前述计算结果，计算截距参数 β_1的95%置信区间为：

$$\begin{aligned} \hat{\beta}_1 - t_{\alpha/2} \cdot S_{\hat{\beta}_1} \quad &\leq \beta_1 \leq \quad \hat{\beta}_1 + t_{\alpha/2} \cdot S_{\hat{\beta}_1} \\ -1.9395 \quad &\leq \beta_1 \quad \leq 1.9106 \end{aligned}$$

- **步骤4**：那么我们可以对截距参数 β_1做出如下检验判断：
  - 不能拒绝原假设 H_0，暂时接受 H_0。认为，长期来看很多个区间[-1.9395,1.9106] 有95%的可能性包含0（ $\beta_1 = 0$）。

假设检验：显著性检验法

显著性检验方法(test-of-significance approach)：是一种用样本结果来证实 H_0真伪的检验程序。

关键思路：

- 找到一个适合的检验统计量(test statistic) 。例如t统计量 χ^2统计量、F统计量等。
- 知道该统计量在 H_0下的抽样分布(pdf)。往往与待检验参数有关系。
- 计算样本统计量的值。也即能用样本数据快速计算出来，例如 $t^*_{\hat{\beta}_2} = \frac{\hat{\beta}_2}{S_{\hat{\beta}_2}}$。
- 查表找出给定显著性水平 α下的**理论统计量**的临界值。例如
  $t_{1-\alpha/2}(n-2) = t_{0.975}(11) = 2.2010$
- 比较样本统计量值和该临界值的大小。例如，比较 $t^*_{\hat{\beta}_2}$ 与 $t_{0.975}(11)$
- 做出拒绝还是接受 H_0的判断。

假设检验：截距参数的t检验

对于截距参数 β_1的显著性检验（t检验）。

- **步骤1**：给出模型，并提出假设：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

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

- **步骤2**：构造合适的检验统计量

$$T = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} \quad \leftarrow T \sim t(n-2)$$

假设检验：截距参数的t检验

- **步骤3**：基于原假设 H_0 计算出样本统计量。

$$T = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} \quad \leftarrow T \sim t(n-2)$$

$$t^*_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} \quad \leftarrow H_0 : \beta_1 = 0$$

$$t^*_{\hat{\beta}_1} = \frac{-0.0145}{0.8746} = -0.0165$$

- **步骤4**：给定显著性水平 $\alpha = 0.05$ 下，查出统计量的**理论分布值**。

> $t_{1-\alpha/2}(n-2) = t_{1-0.05/2}(13-2) = t_{0.975}(11) = 2.2010$

假设检验：截距参数的t检验

- **步骤5**：得到显著性检验的判断结论。
  - 若 $|t^*_{\hat{\beta}_1}| > t_{1-\alpha/2}(n-2)$，则 β_1的t检验结果**显著**。换言之，在显著性水平 $\alpha = 0.05$下，应**显著**地拒绝原假设 H_0，接受备择假设 H_1，认为截距参数 $\beta_1 \neq 0$。
  - 若 $|t^*_{\hat{\beta}_1}| < t_{1-\alpha/2}(n-2)$，则 β_1的t检验结果**不显著**。换言之，在显著性水平 $\alpha = 0.05$下，不能**显著**地拒绝原假设 H_0，只能暂时接受原假设 H_0，认为截距参数 $\beta_1 = 0$。

本例中，$|t^*_{\hat{\beta}_1}| = 0.0165$ 小于 $t_{0.975}(11) = 2.2010$。因此，认为 β_1的t检验结果**不显著**。

换言之，在显著性水平 $\alpha = 0.05$下，不能**显著**地拒绝原假设 H_0，只能暂时接受原假设 H_0，认为截距参数 $\beta_1 = 0$。

假设检验：斜率参数的t检验

对于斜率参数 β_2的显著性检验（t检验）。

- **步骤1**：给出模型，并提出假设：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

$$H_0 : \beta_2 = 0; \quad H_1 : \beta_2 \neq 0$$

- **步骤2**：构造合适的检验统计量

$$T = \frac{\hat{\beta}_2 - \beta_2}{S_{\beta_2}} \quad \leftarrow T \sim t(n-2)$$

假设检验：斜率参数的t检验

- **步骤3**：基于原假设 H_0 计算出样本统计量。

$$T = \frac{\hat{\beta}_2 - \beta_2}{S_{\hat{\beta}_2}} \quad \leftarrow T \sim t(n-2)$$

$$t^*_{\hat{\beta}_2} = \frac{\hat{\beta}_2}{S_{\hat{\beta}_2}} \quad \leftarrow H_0 : \beta_2 = 0$$

$$t^*_{\hat{\beta}_2} = \frac{0.7241}{0.0696} = 10.4064$$

- **步骤4**：给定显著性水平 $\alpha = 0.05$ 下，查出统计量的**理论分布值**。

> $t_{1-\alpha/2}(n-2) = t_{1-0.05/2}(13-2) = t_{0.975}(11) = 2.2010$

假设检验：斜率参数的t检验

- **步骤5**：得到显著性检验的判断结论。
  - 若 $|t^*_{\hat{\beta}_2}| > t_{1-\alpha/2}(n-2)$，则 β_2的t检验结果**显著**。换言之，在显著性水平 $\alpha = 0.05$下，应**显著**地拒绝原假设 H_0，接受备择假设 H_1，认为斜率参数 $\beta_2 \neq 0$。
  - 若 $|t^*_{\hat{\beta}_2}| < t_{1-\alpha/2}(n-2)$，则 β_2的t检验结果**不显著**。换言之，在显著性水平 $\alpha = 0.05$下，不能**显著**地拒绝原假设 H_0，只能暂时接受原假设 H_0，认为斜率参数 $\beta_2 = 0$。

本例中，$|t^*_{\hat{\beta}_2}| = 10.4064$ 大于 $t_{0.975}(11) = 2.2010$。因此，认为 β_2的t检验结果**显著**。

换言之，在显著性水平 $\alpha = 0.05$下，应**显著**地拒绝原假设 H_0，接受备择假设 H_1，认为斜率参数 $\beta_2 \neq 0$。

假设检验：显著性水平VS显著性概率

我们可以回顾犯错误类型：

- 第I类错误：弃真错误 $\alpha = P(Z > Z_0|H_0 = True)$
- 第II类错误：取伪错误 $\beta = P(Z \leq Z_0|H_1 = True)$
- [给定样本容量时]如果我们要减少犯第I 类错误， 第II类错误就要增加；反之亦然。

为什么选择显著性水平 α通常固定在0.01、0.05、0.1水平上？

- 约定而已，并非神圣不可改变！
- 如何改变？？

假设检验：显著性水平VS显著性概率

精确的显著性概率水平p值：

- 对给定的样本算出一个检验统计量(如t统计量)，查到与之对应的概率：p值(p value)或概率值(probability value)
- 不约定α，而是直接求出犯错误概率p值，由读者自己去评判犯错误的可能性和代价！！因人而异！！

假设检验：实际操作中的若干问题

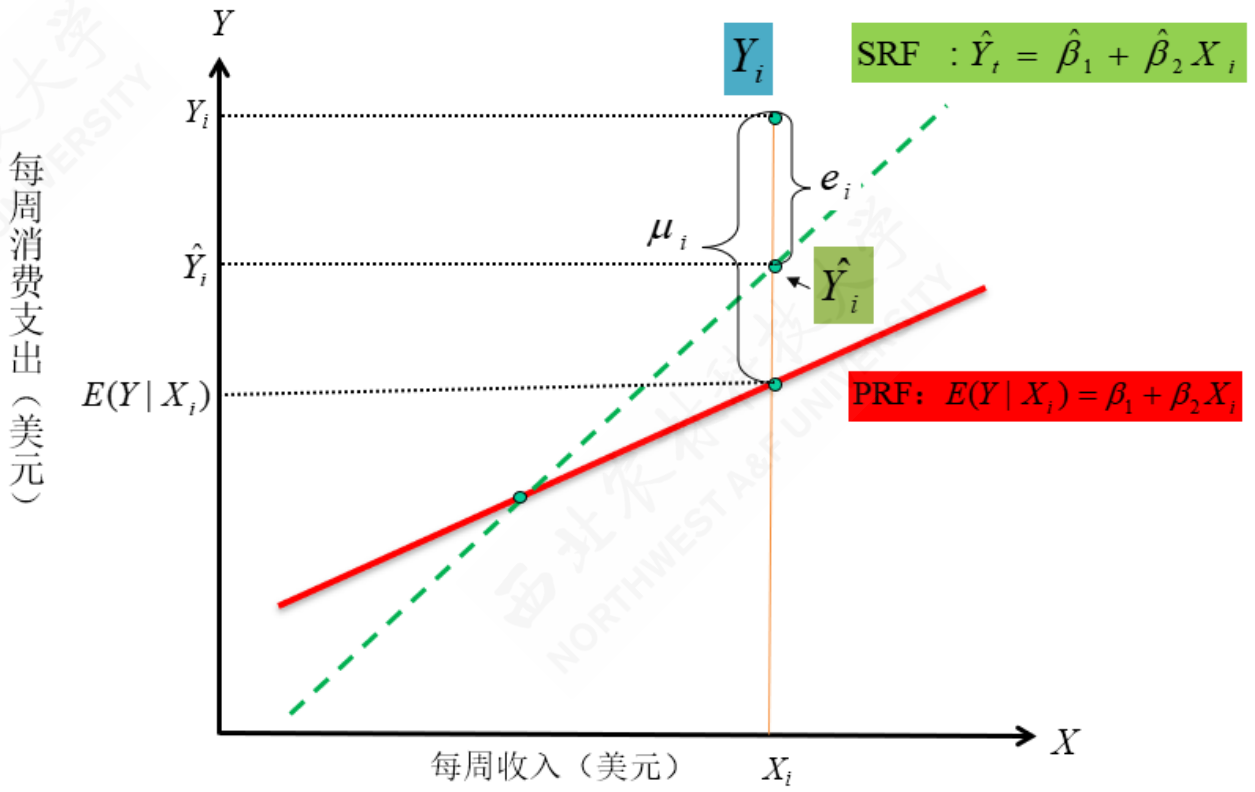
关于**统计显著性**与**实际显著性**。

- 不能一味追求统计显著性，有时候还需要考虑“实际显著性”的现实意义。
- 举例说明：
  - 边际消费倾向(MPC)是指GDP每增加1美元带来消费的增加数；宏观理论表明收入乘数为：1/(1-MPC)。
  - 若MPC的95%置信区间为(0.7129,0.7306)，当样本表明MPC的估计值为$\widehat{MPC} = 0.74$（此时，即乘数为3.84），你怎样抉择！！！

关于**置信区间方法**和**显著性检验方法**的选择。

- 一般来说，置信区间方法优于显著性检验方法！
- 例如：假设MPC $H_0 : \beta_2 = 0$显然荒谬的！

方差分解（ANOVA）：Y变异的分解



$$(Y_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$$

$$y_i = \hat{y}_i + e_i$$

方差分解（ANOVA）：平方和分解

$$
\begin{aligned}
(Y_i - \bar{Y}) &= (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i) \\
y_i &= \hat{y}_i + e_i \\
\sum y_i^2 &= \sum \hat{y}_i^2 + \sum e_i^2 \\
TSS &= ESS + RSS
\end{aligned}
$$

- 其中：TSS表示**总离差平方和**; ESS表示**回归平方和**; RSS表示**残差差平方和**

（附录）：平方和分解证明过程

$$
\begin{aligned}
\sum y_i^2 &= \sum (\hat{y}_i e_i)^2 \\
&= \sum (\hat{y}_i^2 + 2\hat{y}_i e_i + e_i^2) \\
&= \sum \hat{y}_i^2 + 2\sum \hat{y}_i e_i + \sum e_i^2 \\
&= \sum \hat{y}_i^2 + 2\sum \Big((\hat{\beta}_2 x_i) e_i\Big) + \sum e_i^2 \\
&= \sum \hat{y}_i^2 + 2\hat{\beta}_2 \sum (x_i e_i) + \sum e_i^2 \qquad \leftarrow \Big[\sum x_i e_i = 0\Big] \\
&= \sum \hat{y}_i^2 + \sum e_i^2
\end{aligned}
$$

方差分解（ANOVA）：双变量分解表

变异来源	平方和符号SS	平方和计算公式	自由度df	均方和符号MSS	均方和计算公式
回归平方和	ESS	$\sum(\hat{Y}_i-\bar{Y}_i)^2=\sum\hat{y}_i^2$	1	MSS_{ESS}	$ESS/df_{ESS}=\hat{\beta}_2^2\sum x_i^2$
残差平方和	RSS	$\sum(Y_i-\hat{Y}_i)^2=\sum e_i^2$	n-2	MSS_{RSS}	$RSS/df_{RSS}=\frac{\sum e_i^2}{n-2}$
总平方和	TSS	$\sum(Y_i-\bar{Y}_i)^2=\sum y_i^2$	n-1	MSS_{TSS}	$TSS/df_{TSS}=\frac{\sum y_i^2}{n-1}$

模型整体显著性检验：F检验

- **步骤1**：给出模型，并提出假设：

一元回归模型下：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

$$H_0 : \beta_2 = 0; \quad H_1 : \beta_2 \neq 0$$

多元回归模型下：

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i$$

$$H_0 : \beta_2 = \beta_3 = \cdots = \beta_k = 0; \quad H_1 : \text{not all} \quad \beta_j = 0, \quad j \in 2, 3, \cdots, k$$

模型整体显著性检验：F检验

- **步骤2：**构造合适的检验统计量

$$\chi_1^2 = \left(\frac{\hat{\beta}_2 - \beta_2}{\sigma_{\hat{\beta}_2}}\right)^2 = \left(\frac{\hat{\beta}_2 - \beta_2}{\sqrt{\sigma^2 / \sum x_i^2}}\right)^2 = \frac{\left(\hat{\beta}_2 - \beta_2\right)^2 \sum x_i^2}{\sigma^2} \qquad \leftarrow \chi_1^2 \sim \chi^2(1)$$

$$\chi_2^2 = (n-2)\frac{\hat{\sigma}^2}{\sigma^2} = \frac{\sum e_i^2}{\sigma^2} \qquad \leftarrow \chi_2^2 \sim \chi^2(n-2)$$

$$F = \frac{\chi_1^2/1}{\chi_2^2/n-2} = \left(\frac{\left(\hat{\beta}_2 - \beta_2\right)^2 \sum x_i^2}{\sigma^2}\right) / \left(\frac{\sum e_i^2}{(n-2)\sigma^2}\right) = \frac{\left(\hat{\beta}_2 - \beta_2\right)^2 \sum x_i^2}{\sum e_i^2/(n-2)}$$

$$F \sim F(1, n-2)$$

模型整体显著性检验：F检验

- **步骤3**：基于原假设 H_0 计算出样本统计量。

$$
\begin{aligned}
F^* &= \frac{\left(\hat{\beta}_2 - \beta_2\right)^2 \sum x_i^2}{\sum e_i^2/(n-2)} && \leftarrow H_0 : \beta_2 = 0 \\
&= \frac{\hat{\beta}_2^2 \sum x_i^2}{\sum e_i^2/(n-2)} \\
&= \frac{ESS/df_{ESS}}{RSS/df_{RSS}} = \frac{MSS_{ESS}}{MSS_{RSS}} = \frac{\hat{\beta}_2^2 \sum x_i^2}{\hat{\sigma}^2}
\end{aligned}
$$

模型整体显著性检验：F检验

- **步骤4**：给定显著性水平 $\alpha = 0.05$下，查出统计量的**理论分布值**。 $F_{1-\alpha}(1, n-2)$
- **步骤5**：得到显著性检验的判断结论。
  - 若 $F^* > F_{1-\alpha}(1, n-2)$，则 模型整体显著性的F检验结果**显著**。换言之，在显著性水平 $\alpha = 0.05$下，应**显著**地拒绝原假设 H_0，接受备择假设 H_1，认为斜率参数 $\beta_2 \neq 0$。
  - 若 $F^* < F_{1-\alpha}(1, n-2)$，则 模型整体显著性的F检验结果**不显著**。换言之，在显著性水平 $\alpha = 0.05$下，不能**显著**地拒绝原假设 H_0，只能暂时接受原假设 H_0，认为斜率参数 $\beta_2 = 0$。

模型整体显著性检验：比较

F检验与t检验的**联系**：

- 在一元回归模型中，t检验与F检验的结论总是一致的。
- 对于检验斜率参数 β_2的显著性，两者可相互替代！在一元回归分析中，若假设 $H_0: \beta_2 = 0$，则 $F^* \simeq (t^*)^2$

F检验与t检验的**不同**：

- 检验目的不同。F检验是检验模型的整体显著性；t检验是检验各个回归参数的显著性。
- 假设的提出不同：
  - F检验：斜率系数联合假设 $H_0: \beta_2 = 0; \quad H_1: \beta_2 \neq 0$
  - t检验：回归系数分别假设 $H_0: \beta_i = 0; \quad H_1: \beta_i \neq 0; \quad i \in 1, 2$
- 检验原理的不同：F检验需要构造F统计量；t检验需要构造t统计量

（案例）教育程度与时均工资：计算ANOVA表

教育程度与时均工资案例的ANOVA分析表

变异来源	平方和SS	自由度df	均方和MSS
回归平方和ESS	95.4	1	95.42
残差平方和RSS	9.7	11	0.88
总平方和TSS	105.1	12	7.09

（案例）教育程度与时均工资：F检验

- **步骤1**：给出模型 $Y_i = \beta_1 + \beta_2 X_i + u_i$，提出假设：$H_0 : \beta_2 = 0$；　$H_1 : \beta_2 \neq 0$
- **步骤2**：构造合适检验的分布：

$$F = \frac{\left(\hat{\beta}_2 - \beta_2\right)^2 \sum x_i^2}{\sum e_i^2/(n-2)} \qquad \leftarrow F \sim F(1, n-2)$$

- **步骤3**：基于原假设 $H_0 : \beta_2 = 0$，可以计算出样本统计量。

$$F^* = \frac{\hat{\beta}_2^2 \sum x_i^2}{\sum e_i^2/(n-2)} = \frac{ESS/df_{ESS}}{RSS/df_{RSS}} = \frac{MSS_{ESS}}{MSS_{RSS}} = \frac{95.4253}{0.8812} = 108.2924$$

- **步骤4**：给定 $\alpha = 0.05$下，查出F**理论值** $F_{1-\alpha}(1, n-2) = F_{0.95}(1, 11) = 4.8443$
- **步骤5**：得到显著性检验的判断结论。因为 $F^* = 108.2924$ 大于 $F_{0.95}(1, 11) = 4.8443$，所以模型整体显著性的F检验结果**显著**。换言之，在显著性水平 $\alpha = 0.05$ 下，应**显著**地拒绝原假设 H_0，接受备择假设 H_1，认为斜率参数 $\beta_2 \neq 0$。

5.5 拟合优度与残存分析

拟合优度

残存分析

拟合优度：引子

怎么来判定OLS方法对特定样本数据拟合的好坏？

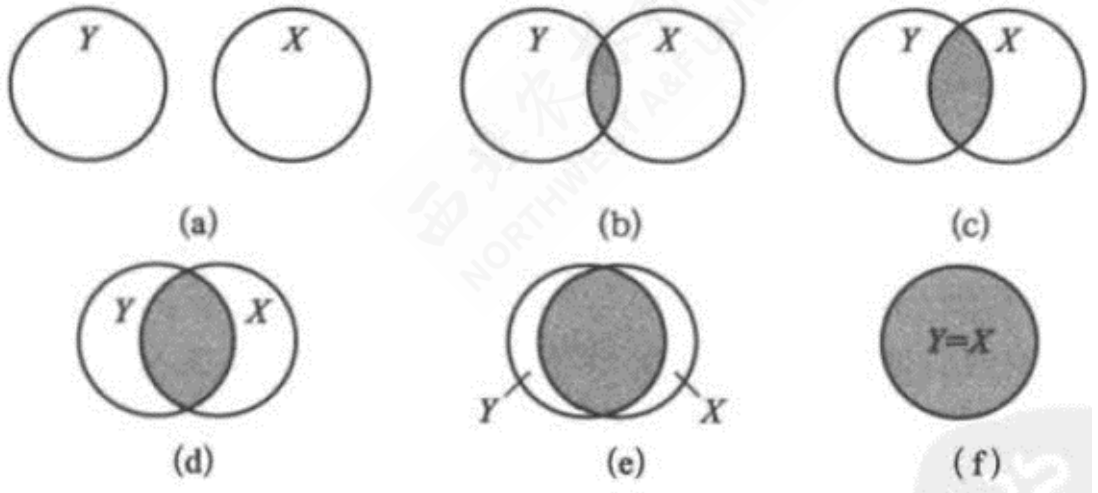
请大家思考如下几个**问题**：

- 样本数据不完全落在拟合的直线（或曲线）上，是经常发生的么？
- 怎么来表达或测量这种对样本数据拟合的不完全性？
- 在OLS方法和CLRM假设“双剑合璧”下，对特定样本数据的拟合不是已经证明最好的么（BLUE）？为什么还要说“拟合”有“好坏之分”？

拟合优度：测量指标

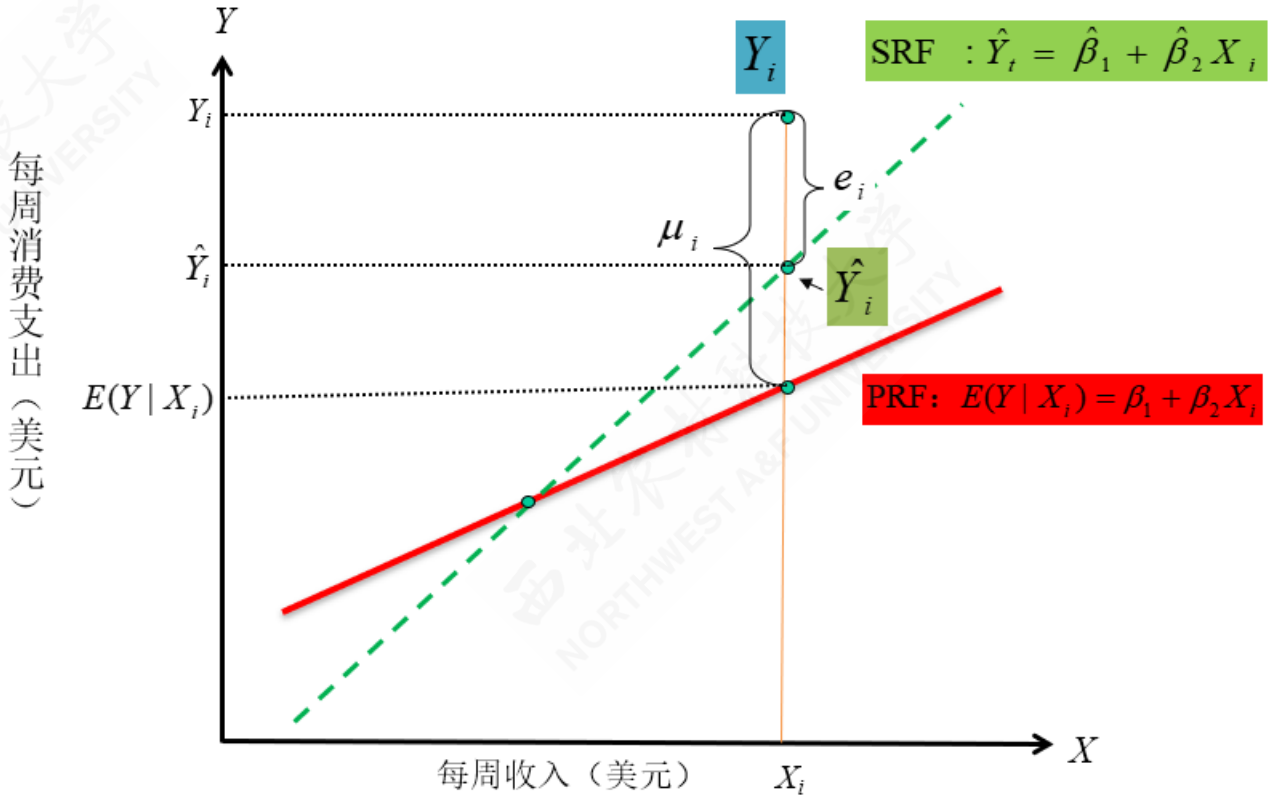
拟合优度（Goodness of fit）：判断样本回归线对一组数据拟合优劣水平的度量。

判定系数（coefficient of determination）：一种利用平方和分解，考察样本回归线对数据拟合效果的总度量。一元回归中，一般记为 r^2；多元回归中，一般记为 R^2。



维恩图看拟合优度

拟合优度：测量指标



平方和分解看拟合优度

拟合优度：判定系数

判定系数r^2计算公式1：

$$r^2 = \frac{ESS}{TSS} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$$

判定系数r^2计算公式2：

$$r^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum(Y_i - \bar{Y})^2}$$

拟合优度：判定系数

判定系数r^2计算公式3：

$$r^2 = \frac{ESS}{TSS} = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = \frac{\sum (\hat{\beta}_2 x_i)^2}{\sum y_i^2} = \hat{\beta}_2^2 \frac{\sum x_i^2}{\sum y_i^2} = \hat{\beta}_2^2 \frac{S_{X_i}^2}{S_{Y_i}^2}$$

判定系数r^2计算公式4：

$$r^2 = \hat{\beta}_2^2 \cdot \frac{\sum x_i^2}{\sum y_i^2} = \left(\frac{\sum x_i y_i}{\sum x_i^2}\right)^2 \cdot \left(\frac{\sum x_i^2}{\sum y_i^2}\right) = \frac{(\sum x_i y_i)^2}{\sum x_i^2 \sum y_i^2}$$

课堂讨论：

- 讨论1：r^2是一个非负量。为什么？
- 讨论2：$0 \leq r^2 \leq 1$，两个端值分别意味什么？

拟合优度：判定系数VS简单相关系数

判定系数与简单相关系数有什么区别与联系？

总体相关系数：是变量 X_i与变量 Y_i总体相关关系的参数，一般记为 ρ。

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X_i)Var(Y_i)}} = \frac{E(X_i - EX)(Y_i - EY)}{\sqrt{E(X_i - EX)^2 E(Y_i - EY)^2}}$$

样本相关系数：是从总体中抽取随机样本，获得变量 X_i与变量 Y_i样本相关关系的统计量度量，一般记为 r。

$$r = \frac{S_{XY}^2}{S_X * S_Y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$$

拟合优度：判定系数VS简单相关系数

判定系数和简单相关系数的联系：

- 在一元回归中，判定系数r^2等于样本相关系数r的平方。

判定系数和简单相关系数的区别：

- 判定系数r^2表明因变量变异由解释变量所解释的比例，而相关系数r只能表明变量间的线性关联强度。
- 在多元回归中，这种区别会更加凸显！因为那时的相关系数r出现了偏相关的情形(交互关联)！

（案例）计算相关系数和判定系数

对于“教育程度与时均工资案例”，根据FF-ff计算表和方差分解ANOVA表，可以分别计算得到样本相关系数和模型判定系数。

样本相关系数 r：

$$r = \frac{S_{XY}^2}{S_X * S_Y} = \frac{10.9821}{3.8944 * 2.9597} = 0.9528$$

回归方程的判定系数 r^2：

$$r^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{9.693}{105.1183} = 0.9078$$

二者关系

拟合优度：小结与思考

内容小结：

- 即使采用OLS方法，它对样本数据的拟合也是不完全的。意味着实际数据点在样本回归线附近，而不是在样本回归线上。我们可以把样本点行为的“变异”，划分为“回归”能解释的部分和“随机”的部分。并进一步获得变异平方和的分解。
- 判定系数 R^2是对OLS拟合程度的测量，它使用了变异平方和分解的思想。在一元线性回归（含截距）中，判定系数与相关系数存在如下关系 $R^2 = r^2_{(X_i,Y_i)}$。注意，在多元回归中则不存在这种关系。

问题思考：

- OLS方法的参数估计量，在CLRM假设满足情况下，就是最优线性无偏估计量（BLUE），为什么还要用**判定系数**来判断“拟合好还是不好？”。对此，你的回答是什么？
- 还有没有其他指标，来反映估计方法对样本数据的拟合好坏程度？请说出一两个。

残差分析：定义和作用

残差(residual)：是因变量的观测值与根据估计的回归方程求出的估计值之差，用e_i表示。

$$e_i = Y_i - \hat{Y}_i$$

对模型的残差进行分析，主要目的包括：

- 反映用估计的回归方程去预测而引起的误差。
- 可用于确定有关随机干扰项μ_i的假定是否成立。
- 用于检测有影响的观测值。

残差分析：皮尔逊标准化残差

标准化残差(standardized residual)：是对残差进行某种标准化变换。具体计算方法有**皮尔逊标准化残差**和**学生化标准残差**两种。

最常用的皮尔逊标准化残差（Pearson residual/internally studentized residuals）的计算公式如下：

$$e^*_{i,sd} = \frac{e_i}{s_{e_i}} = \frac{(Y_i - \hat{Y}_i)}{\sqrt{\frac{\sum (e_i - \bar{e})^2}{n-1}}}$$

残差分析：皮尔逊标准化残差

学生化标准残差（Studentized Residuals/externally studentized residual/deleted Studentized residual/semi-studentized residuals/jackknifed residuals），是对残差的另一种特殊标准化变换（例如考虑到了X的影响力）。

学生化标准残差的计算公式有两个*：

$$e_{i,st}^{*}=\frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}}$$

$$e_{i,st}^{*}=e_{i,sd}^{*}\left(\frac{n-m-2}{n-m-1-e_{i,sd}^{*2}}\right)^{2}$$

> 其中：$MSE_{(i)}$是指删除第i个观测值进行建模的**均方误差**（MSE）；h_{ii}指删除第i个观测值进行建模的第i个**影响权重**（leverage）。$m=k-1$为回归元个数。

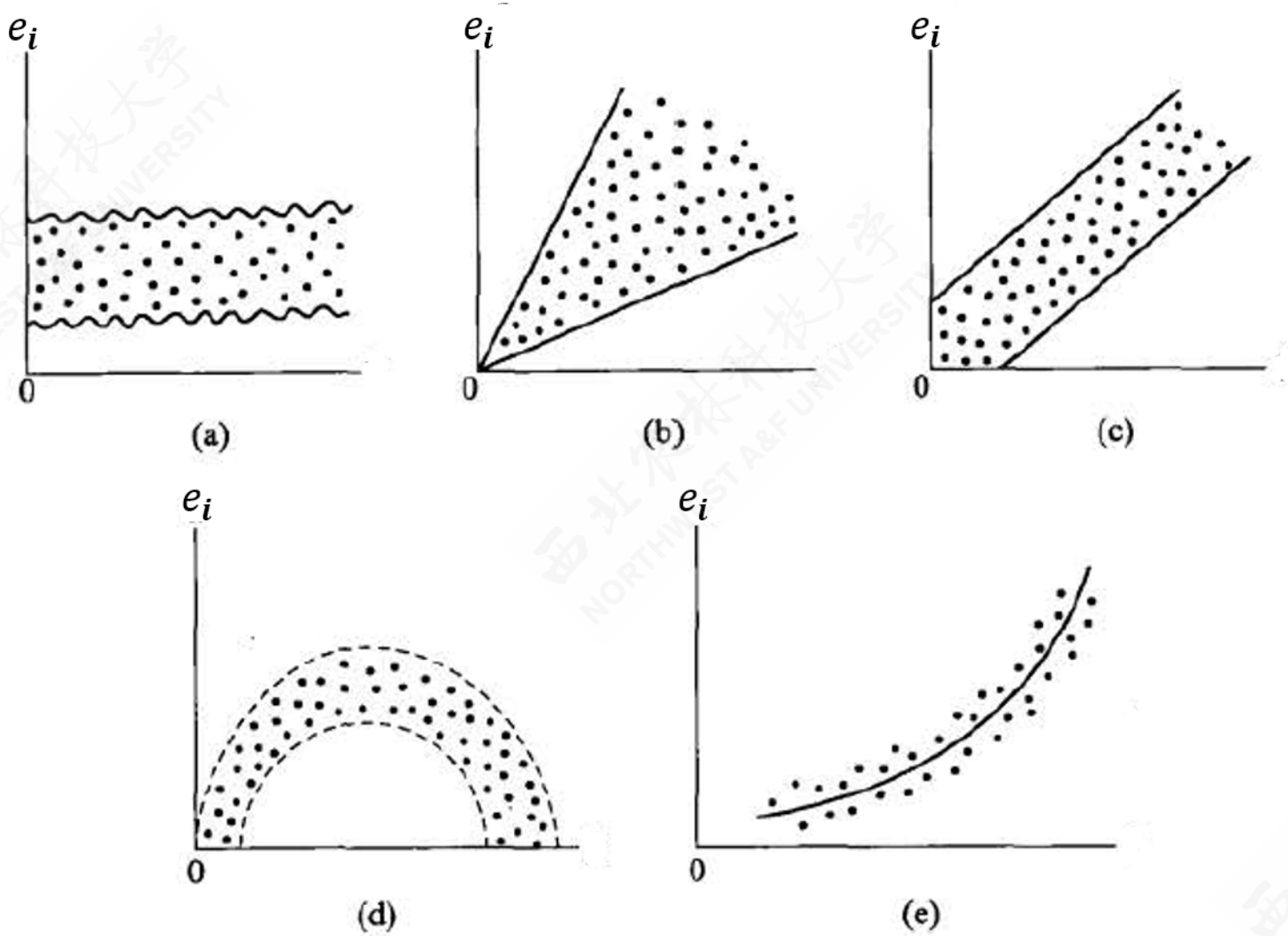
说明：1）学生化残差的第一个计算公式计算起来比较麻烦和复杂。需要分别进行(n-1)次线性回归，然后依次计算相关$MSE_{(i)}$和h_ii。2）学生化残差的第二个计算公式相对简单，只需要利用原来的回归模型及其标准化残差$e_{i,sd}^{*}$。

残差分析：残差图

残差图(residual plot)：用于呈现残差数据 e_i的分布情况的统计图图形，主要包括：

- 关于 X_i的残差散点图。
- 关于 Y_i的残差散点图（或者关于 $\hat{Y}_i$）。
- 关于样本序号的残差散点图或标准化残差散点图。

（示例）残差图的模拟演示



（案例）皮尔逊标准化残差

obs	`$X_i$`	`$Y_i$`	`$\hat{Y}_i$`	`$e_i$`	`$e^*_{i,sd}$`
1	6	4.5	4.3	0.1266	0.1408
2	7	5.8	5.0	0.7158	0.7964
3	8	6.0	5.8	0.2004	0.2230
4	9	7.3	6.5	0.8293	0.9227
5	10	7.3	7.2	0.0917	0.1020
6	11	6.6	8.0	-1.3662	-1.5201
7	12	7.8	8.7	-0.8565	-0.9530
8	13	7.8	9.4	-1.5637	-1.7399
9	14	11.0	10.1	0.8994	1.0007
10	15	10.7	10.8	-0.1732	-0.1927
11	16	10.8	11.6	-0.7350	-0.8178
12	17	13.6	12.3	1.3198	1.4685
13	18	13.5	13.0	0.5117	0.5694
sum	156	112.8	112.8	0.0000	-0.0000

- 根据样本回归方程，可以计算得到 Y_i 的回归拟合值 $\hat{Y}_i$，以及回归残差 e_i。

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$
$$e_i = Y_i - \hat{Y}_i$$

- 进一步地计算得到**皮尔逊标准化残差** $e^*_{i,sd}$：

$$e^*_{i,sd} = \frac{e_i}{s_{e_i}} = \frac{(Y_i - \hat{Y}_i)}{\sqrt{\frac{\sum (e_i - \bar{e})^2}{n-1}}}$$

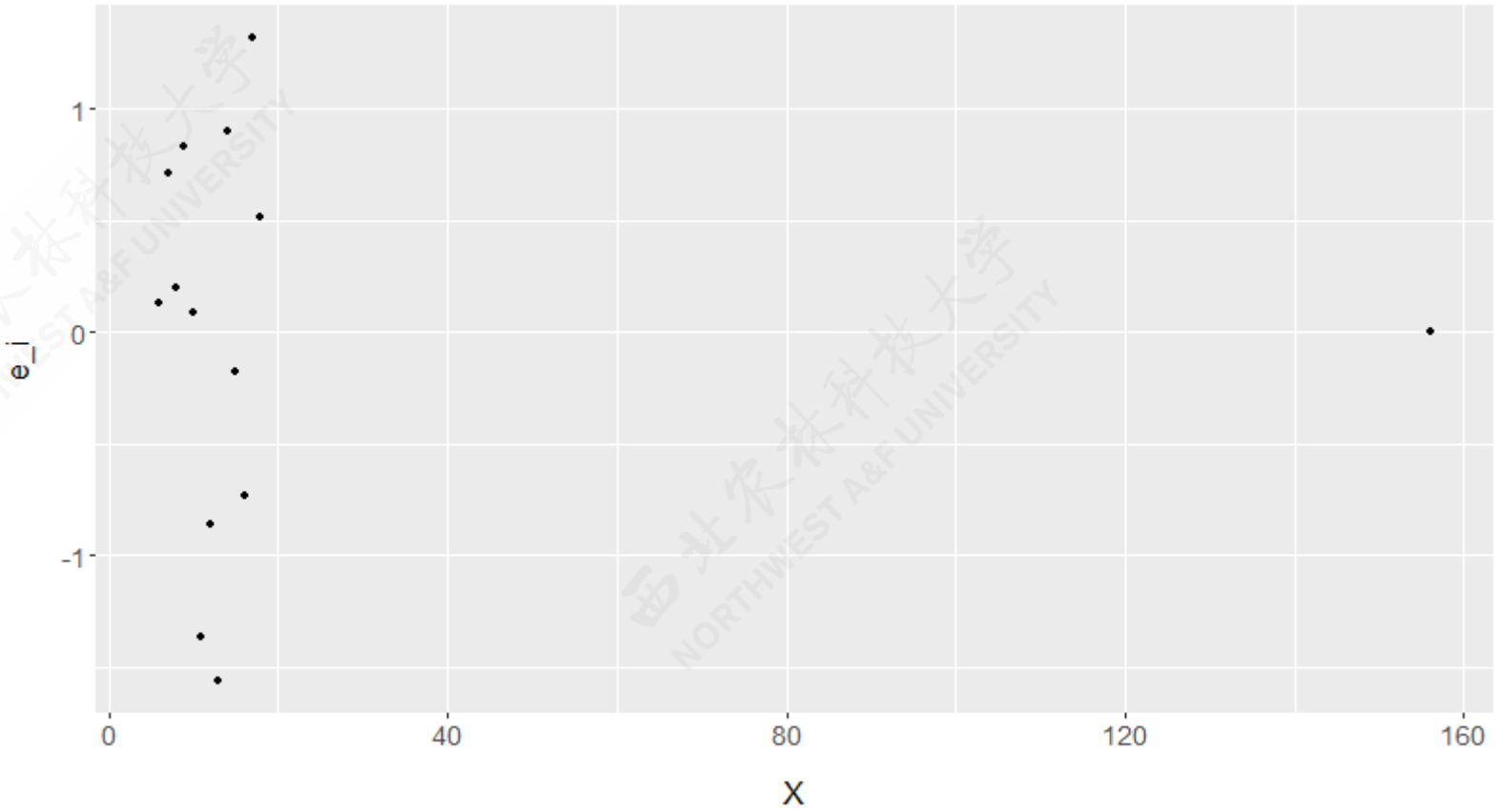
（案例）学生化标准残差

obs	`$X_i$`	`$Y_i$`	`$\hat{Y}_i$`	`$e_i$`	`$e^*_{i,sd}$`	`$e^*_{i,st}$`
1	6	4.5	4.3	0.1266	0.1408	0.1511
2	7	5.8	5.0	0.7158	0.7964	0.8493
3	8	6.0	5.8	0.2004	0.2230	0.2233
4	9	7.3	6.5	0.8293	0.9227	0.9402
5	10	7.3	7.2	0.0917	0.1020	0.0982
6	11	6.6	8.0	-1.3662	-1.5201	-1.6297
7	12	7.8	8.7	-0.8565	-0.9530	-0.9451
8	13	7.8	9.4	-1.5637	-1.7399	-1.9472
9	14	11.0	10.1	0.8994	1.0007	1.0103
10	15	10.7	10.8	-0.1732	-0.1927	-0.1885
11	16	10.8	11.6	-0.7350	-0.8178	-0.8456
12	17	13.6	12.3	1.3198	1.4685	1.7221
13	18	13.5	13.0	0.5117	0.5694	0.6220
sum	156	112.8	112.8	0.0000	-0.0000	0.0601

- 根据样本回归方程，可以计算得到 Y_i 的回归拟合值 $\hat{Y}_i$，以及回归残差 e_i，以及前述的**皮尔逊标准化残差** $e^*_{i,sd}$。
- 进而可以使用如下公式计算得到**学生化标准残差** $e^*_{i,st}$：

$$e^*_{i,st} = e^*_{i,sd}\left(\frac{n-k-2}{n-k-1-e^{*2}_{i,sd}}\right)^2$$

（案例）皮尔逊标准化残差散点图



5.6 回归预测分析

均值预测

个值预测

置信带

回归预测：引子

预测未来事件的一些惯常说法

- 算命术士：
  - “客官印堂发黑，明日必有凶象!”
- 天气预报播报词：
  - 预测西安明天是小雨，概率为95%。
  - 预测西安明天是小雨转阴，概率为95%。
  - 预测西安明天是天晴或阴天或雨天，概率为100%!
- 简要解析：
  - 人们在预测什么事件？
  - 预测多少个事件？它们发生的关系？
  - 预测如何令人信服？

回归预测：两类预测

一元回归模型下：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

预测什么？

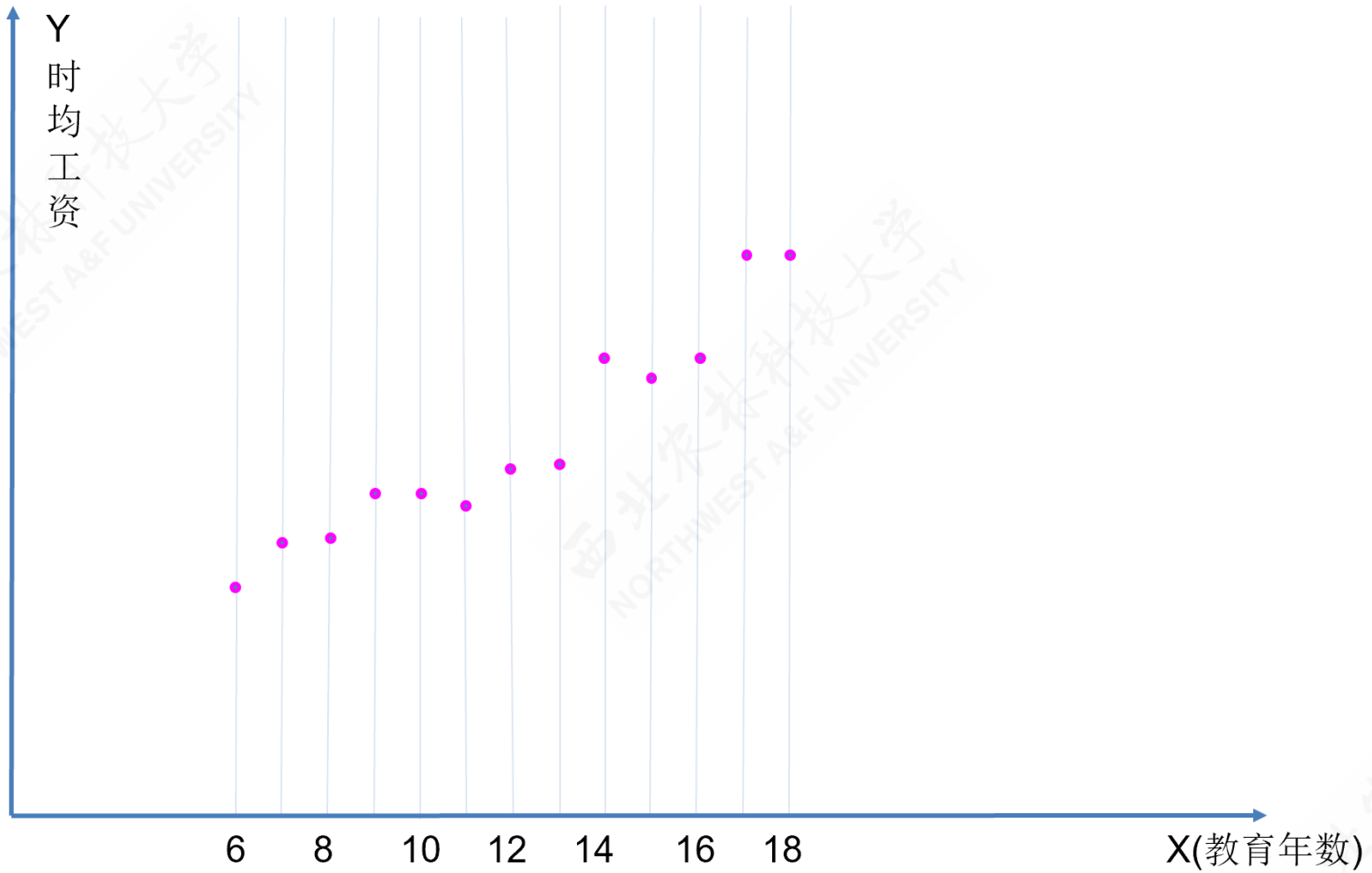
均值预测(mean prediction)：

- 给定 X_0，预测Y的条件均值 $E(Y|X = X_0)$

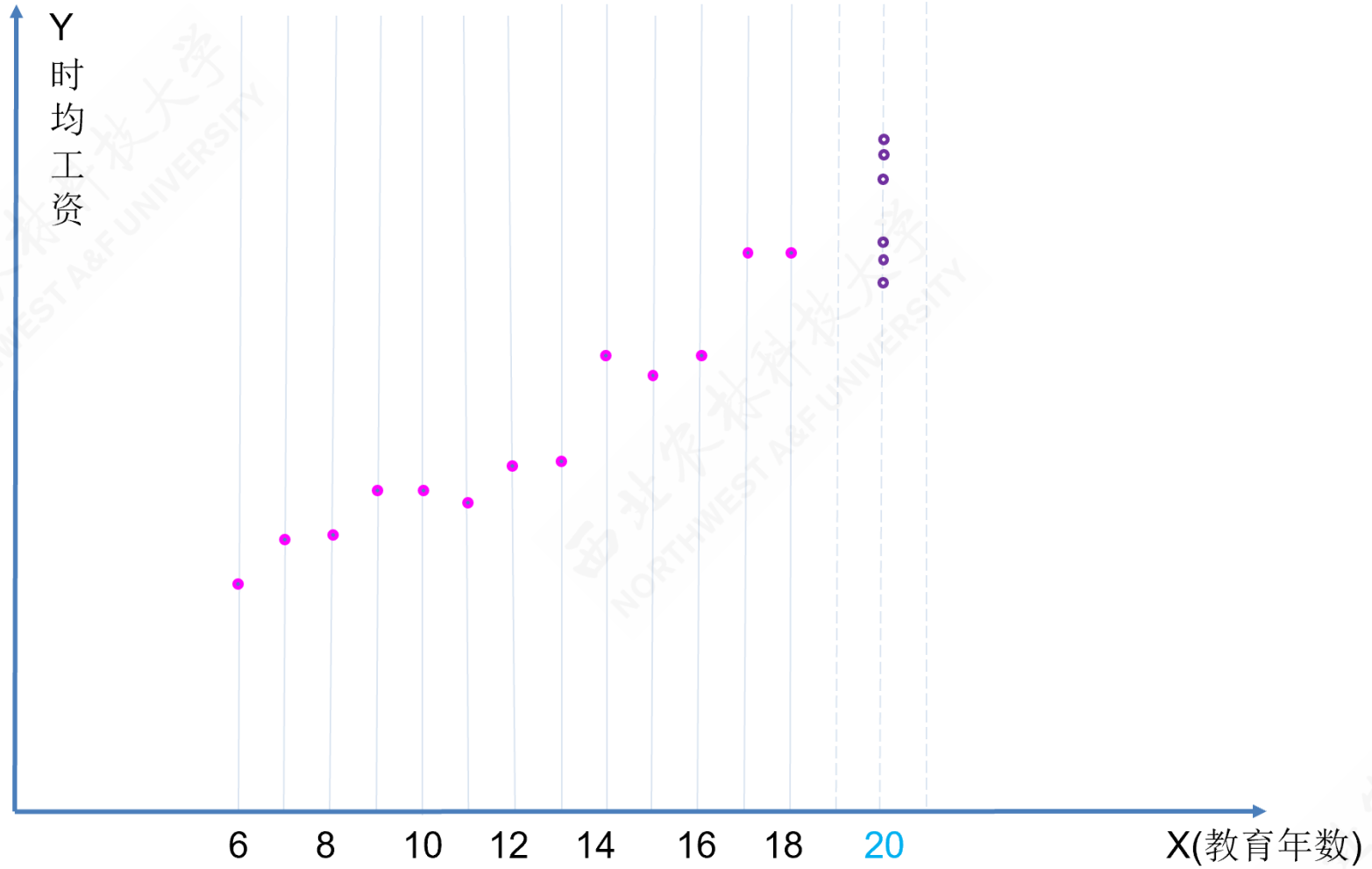
个值预测(individual prediction)：

- 给定 X_0，预测对应于 $X0$的Y的个别值 $(Y_0|X_0)$

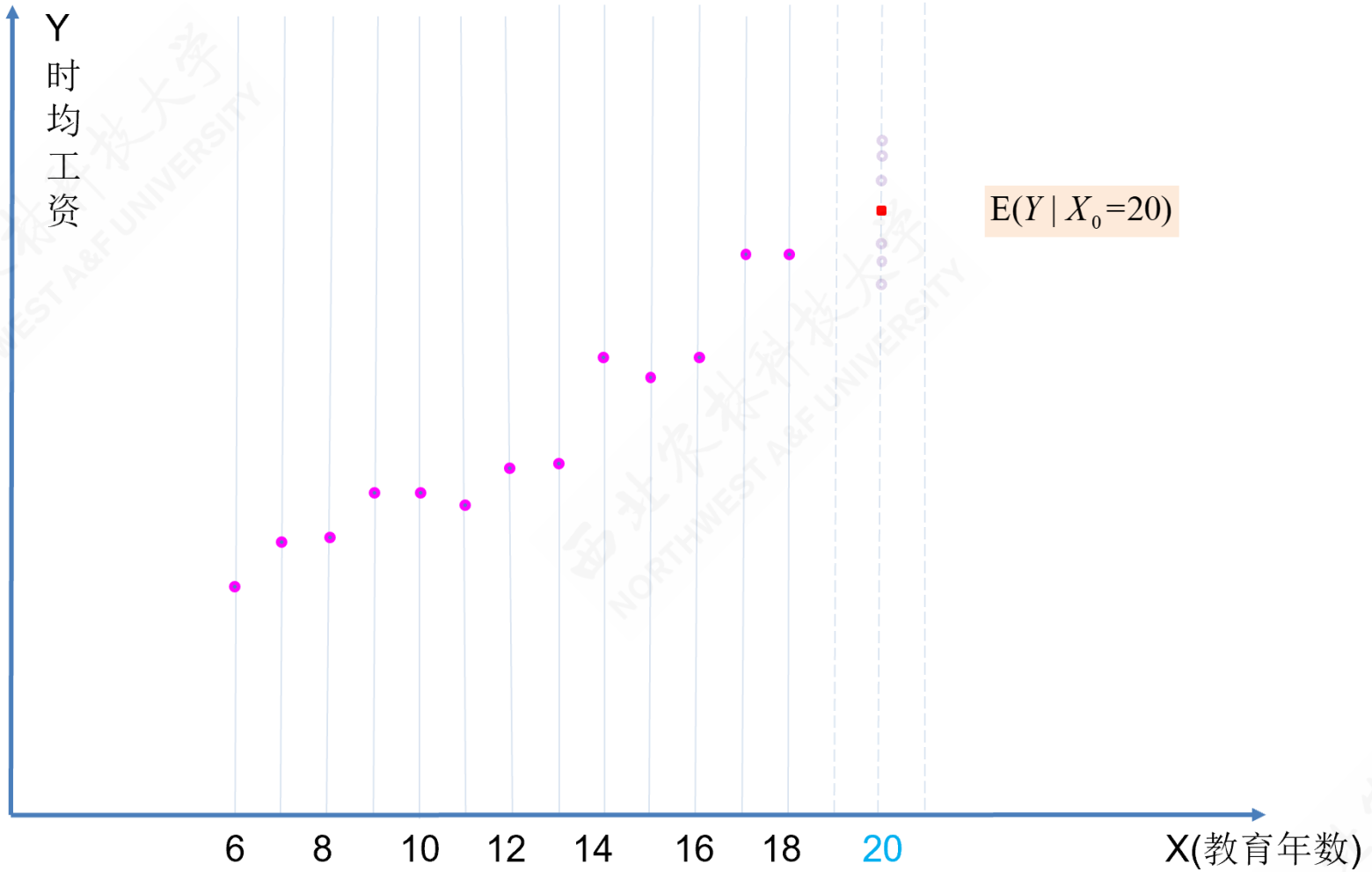
（示例）样本内预测



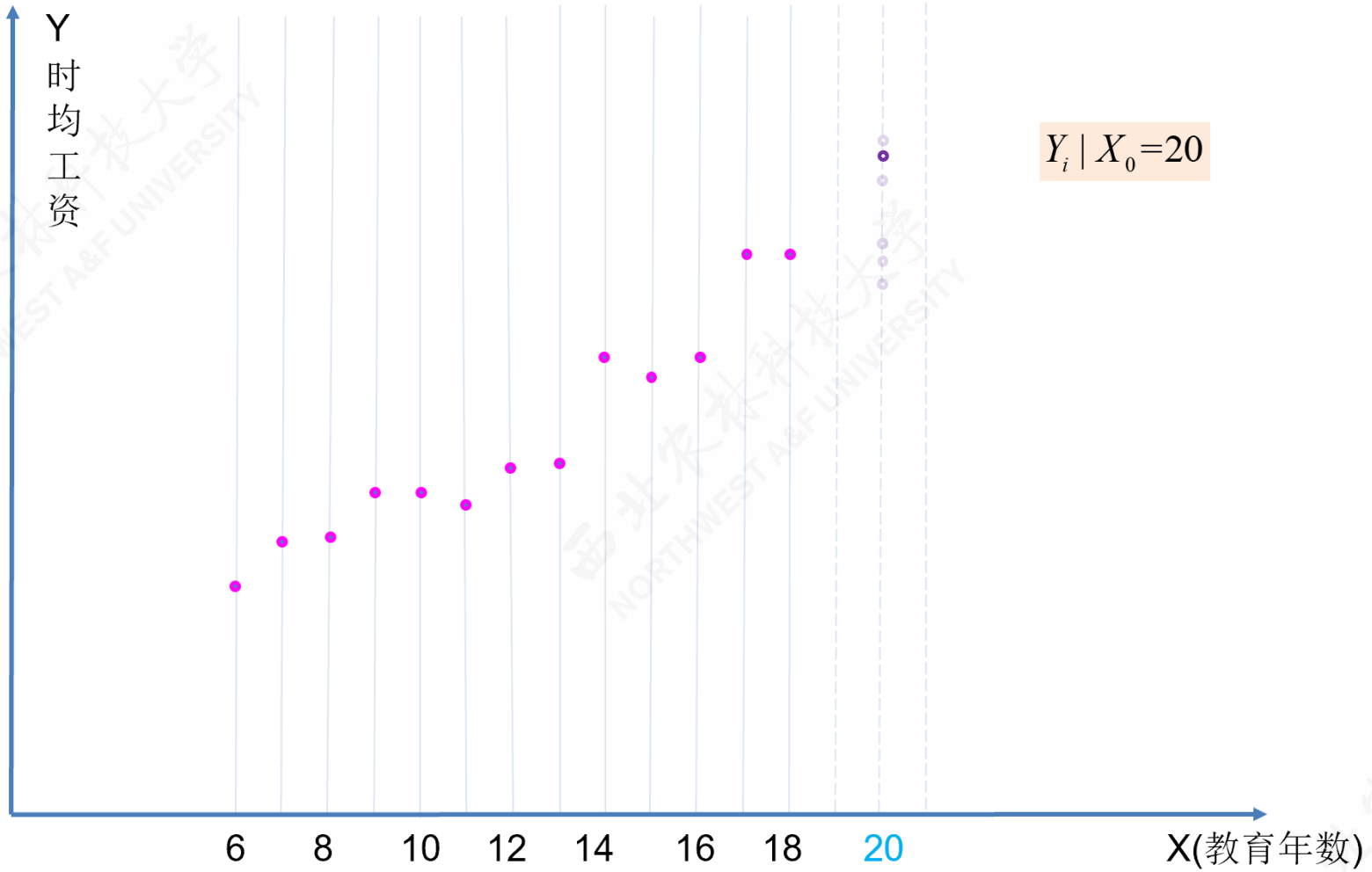
（示例）样本外预测



（示例）均值预测



（示例）个值预测



回归预测：预测分析的关键

拿什么来预测？——样本数据？样本回归线？样本拟合值？

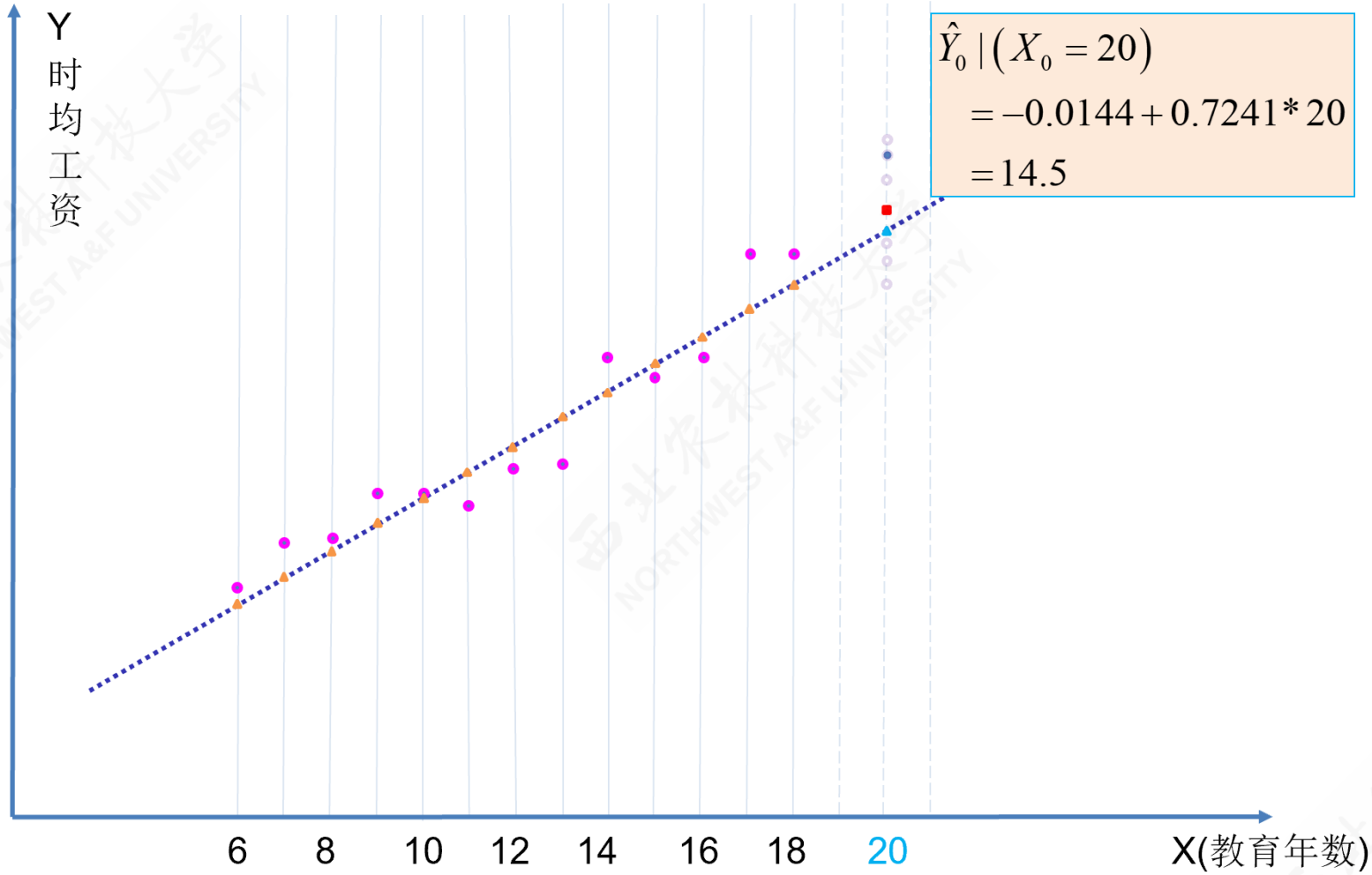
样本外拟合值 $\hat{Y}_0|X=X_0$：

- 可以证明：样本外拟合值 $\hat{Y}_0|X=X_0$是**均值** $E(Y|X=X_0)$的一个**BLUE**
- 也可以证明：样本外拟合值 $\hat{Y}_0|X=X_0$是**个值** $(Y_0|X=X_0)$的一个**BLUE**

工资案例中，给定 $X_0=20$，则可以得到样本外拟合值：

$$\hat{Y}_0=\hat{\beta}_1+\hat{\beta}_2X_0=-0.0145+0.7241*20=14.4675$$

回归预测：预测分析的关键



均值预测

在**N-CLRM**假设和**OLS**方法下，可以证明（证明过程略）给定 X_0 下的拟合值 $\hat{Y}_0$ 服从如下正态分布：

$$\hat{Y}_0 \sim \mathrm{N}\left(\mu_{\hat{Y}_0}, \sigma^2_{\hat{Y}_0}\right)$$

$$\mu_{\hat{Y}_0} = E\left(\hat{Y}_0\right) = E\left(\hat{\beta}_1 + \hat{\beta}_2 X_0\right) = \beta_1 + \beta_2 X_0 = E(Y|X_0)$$

$$\mathrm{var}\left(\hat{Y}_0\right) = \sigma^2_{\hat{Y}_0} = \sigma^2 \left[\frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{\sum x_i^2}\right]$$

$$\hat{Y}_0 \sim N\left(E(Y|X_0), \sigma^2 \left[\frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{\sum x_i^2}\right]\right)$$

均值预测

对 $\hat{Y}_0$构造**t统计量**：

$$T=\frac{\hat{Y}_0-\mathrm{E}(\mathrm{Y}|\mathrm{X}_0)}{S_{\hat{Y}_0}}\sim t(n-2) \qquad \Leftarrow S_{\hat{Y}_0}=\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\left(X_0-\overline{X}\right)^2}{\sum x_i^2}\right]}$$

得到**均值** $E(Y|X=X_0)$置信区间为：

$$\Pr\left[\hat{Y}_0-t_{1-\alpha/2}(n-2)\cdot S_{\hat{Y}_0}\le E(Y|X_0)\le \hat{Y}_0+t_{1-\alpha/2}(n-2)\cdot S_{\hat{Y}_0}\right]=1-\alpha$$

$$\Pr\left[\hat{\beta}+\hat{\beta}_2X_0-t_{1-\alpha/2}(n-2)\cdot S_{\hat{Y}_0}\le E(Y|X_0)\le \hat{\beta}+\hat{\beta}_2X_0+t_{1-\alpha/2}(n-2)\cdot S_{\hat{Y}_0}\right]=1-\alpha$$

给定 $X_0 = 20$时，根据早前计算结果：$\hat{\sigma}^2 = 0.8812$；$\bar{X} = 12.0000$；$\sum x_i^2 = 182.0000$。因此可以得到：

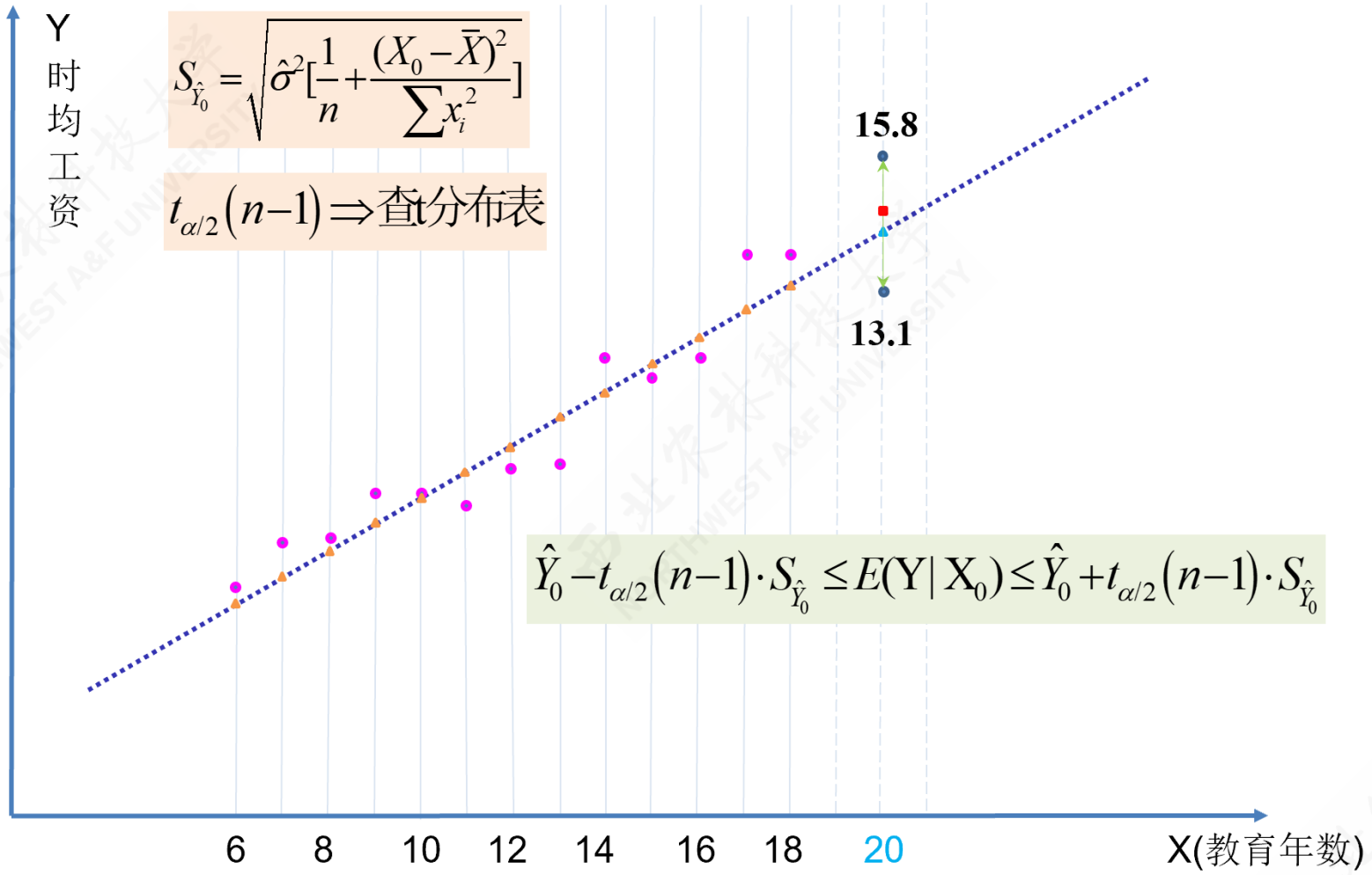
$$S_{\hat{Y}_0}^2 = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{\sum x_i^2} \right] = 0.8812 \left(\frac{1}{13} + \frac{(20-12)^2}{182} \right) = 0.3776; \quad S_{\hat{Y}_0} = \sqrt{S_{\hat{Y}_0}^2} = 0.6145$$

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0 = -0.0145 + 0.7241 * 20 = 14.4675$$

因此，可以计算得到**均值** $E(Y|X=20)$置信区间为：

$$\begin{aligned} \hat{\beta} + \hat{\beta}_2 X_0 - t_{1-\alpha/2}(n-2) \cdot S_{\hat{Y}_0} \leq & E(Y|X_0) \leq \hat{\beta} + \hat{\beta}_2 X_0 + t_{1-\alpha/2}(n-2) \cdot S_{\hat{Y}_0} \\ 14.4675 - 1.7959 * 0.6145 \leq & E(Y|X_0 = 20) \leq 14.4675 + 1.7959 * 0.6145 \\ 13.3639 \leq & E(Y|X_0 = 20) \leq 15.5711 \end{aligned}$$

（案例）教育程度和时均工资：均值预测



个值预测

在**N-CLRM**假设和**OLS**方法下，可以证明（证明过程略）给定 X_0 下的个别值 $Y_0 = \beta_1 + \beta_2 X_0 + u_0$ 服从如下正态分布：

$$
\begin{aligned}
Y_0 &\sim \mathrm{N}\left(\mu_{Y_0}, \sigma^2_{Y_0}\right) \\
\mu_{Y_0} &= E\left(Y_0\right) = E\left(\beta_1 + \beta_2 X_0\right) = \beta_1 + \beta_2 X_0 \\
Var(Y_0) &= Var(u_0) = \sigma^2
\end{aligned}
$$

$$
Y_0 \sim N\left(\beta_1 + \beta_2 X_0, \sigma^2\right)
$$

个值预测

进一步可以构造新的随机变量 $(Y_0 - \hat{Y}_0)$，其将服从如下正态分布：

$$Y_0 \sim N\left(\beta_1 + \beta_2 X_0, \sigma^2\right)$$

$$\hat{Y}_0 \sim N\left(\beta_1 + \beta_2 X_0, \sigma^2\left[\frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{\sum x_i^2}\right]\right)$$

$$Y_0 - \hat{Y}_0 \sim N\left(0, \sigma^2\left[1 + \frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{\sum x_i^2}\right]\right)$$

$$Y_0 - \hat{Y}_0 \sim N\left(0, \sigma^2_{Y_0 - \hat{Y}_0}\right)$$

个值预测

对 $Y_0 - \hat{Y}_0$构造**t统计量**：

$$T = \frac{(Y_0 - \hat{Y}_0)}{S_{(Y_0 - \hat{Y}_0)}} \sim t(n-2) \qquad \Leftarrow S_{(Y_0 - \hat{Y}_0)} = \sqrt{\hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{\sum x_i^2}\right]}$$

得到**个值** Y_0置信区间为：

$$\Pr\left[\hat{Y}_0 - t_{1-\alpha/2}(n-2) \cdot S_{(Y_0 - \hat{Y}_0)} \le Y_0 \le \hat{Y}_0 + t_{1-\alpha/2}(n-2) \cdot S_{(Y_0 - \hat{Y}_0)}\right] = 1 - \alpha$$

$$\Pr\left[\hat{\beta} + \hat{\beta}_2 X_0 - t_{1-\alpha/2}(n-2) \cdot S_{(Y_0 - \hat{Y}_0)} \le Y_0 \le \hat{\beta} + \hat{\beta}_2 X_0 + t_{1-\alpha/2}(n-2) \cdot S_{(Y_0 - \hat{Y}_0)}\right] = 1 - \alpha$$

给定 $X_0 = 20$时，根据早前计算结果：$\hat{\sigma}^2 = 0.8812$；$\bar{X} = 12.0000$；$\sum x_i^2 = 182.0000$。因此可以得到：

$$S^2_{(Y_0-\hat{Y}_0)} = \hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{\sum x_i^2}\right] = 0.8812 \left(1 + \frac{1}{13} + \frac{(20-12)^2}{182}\right) = 1.2588$$

$$S_{\hat{Y}_0} = \sqrt{S^2_{\hat{Y}_0}} = 1.122$$

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0 = -0.0145 + 0.7241 * 20 = 14.4675$$

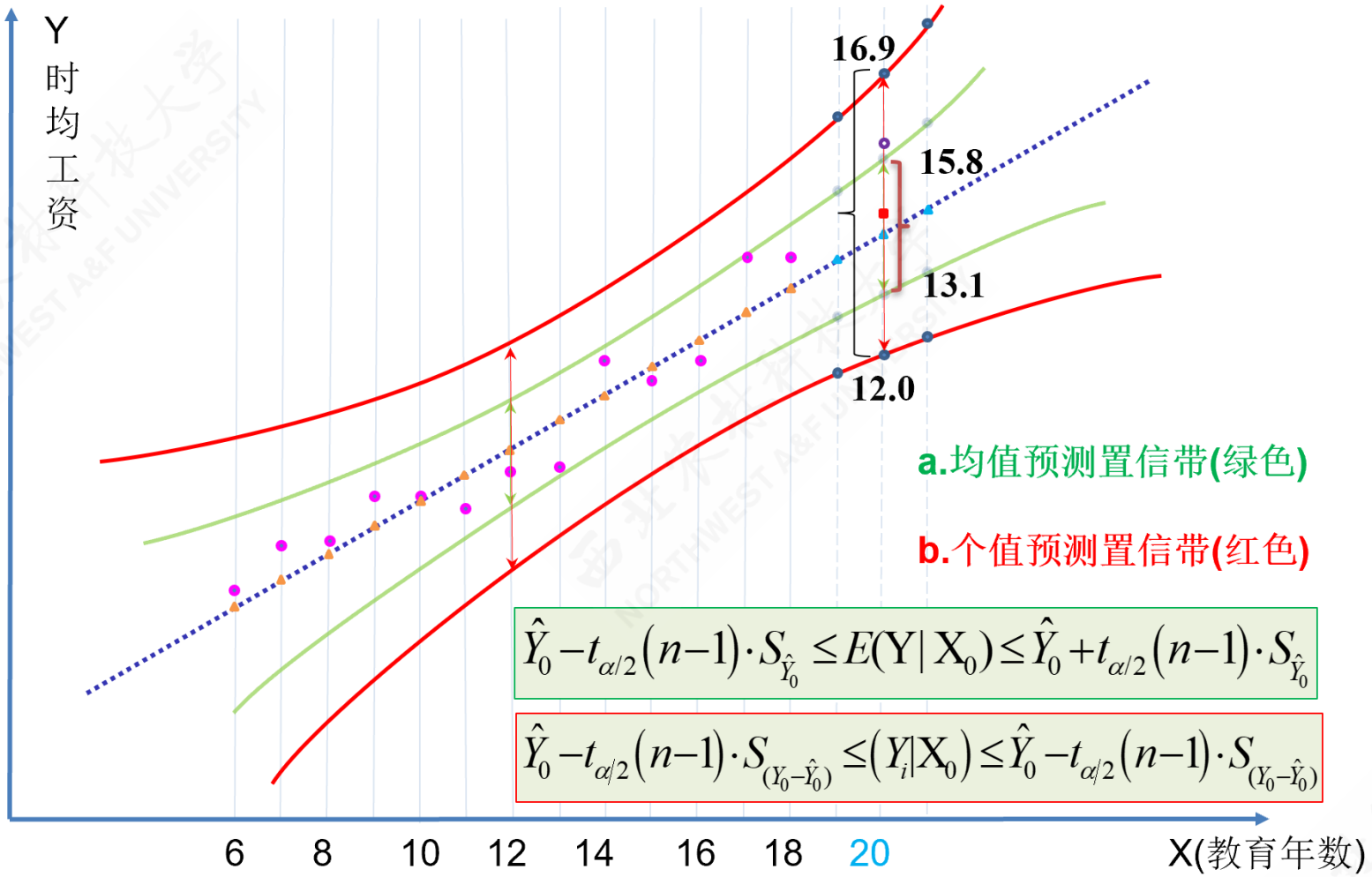
因此，可以计算得到**个值** $(Y_0|X = 20)$置信区间为：

$$\hat{\beta} + \hat{\beta}_2 X_0 - t_{1-\alpha/2}(n-2) \cdot S_{(Y_0-\hat{Y}_0)} \leq Y_0|X = X_0) \leq \hat{\beta} + \hat{\beta}_2 X_0 + t_{1-\alpha/2}(n-2) \cdot S_{(Y_0-\hat{Y}_0)}$$

$$14.4675 - 1.7959 * 1.122 \leq Y_0|X_0 = 20) \leq 14.4675 + 1.7959 * 1.122$$

$$12.4525 \leq Y_0|X_0 = 20) \leq 16.4824$$

（案例）教育程度和时均工资：个值预测



a.均值预测置信带(绿色)

b.个值预测置信带(红色)

$$\hat{Y}_0 - t_{\alpha/2}(n-1)\cdot S_{\hat{Y}_0} \le E(Y|X_0) \le \hat{Y}_0 + t_{\alpha/2}(n-1)\cdot S_{\hat{Y}_0}$$

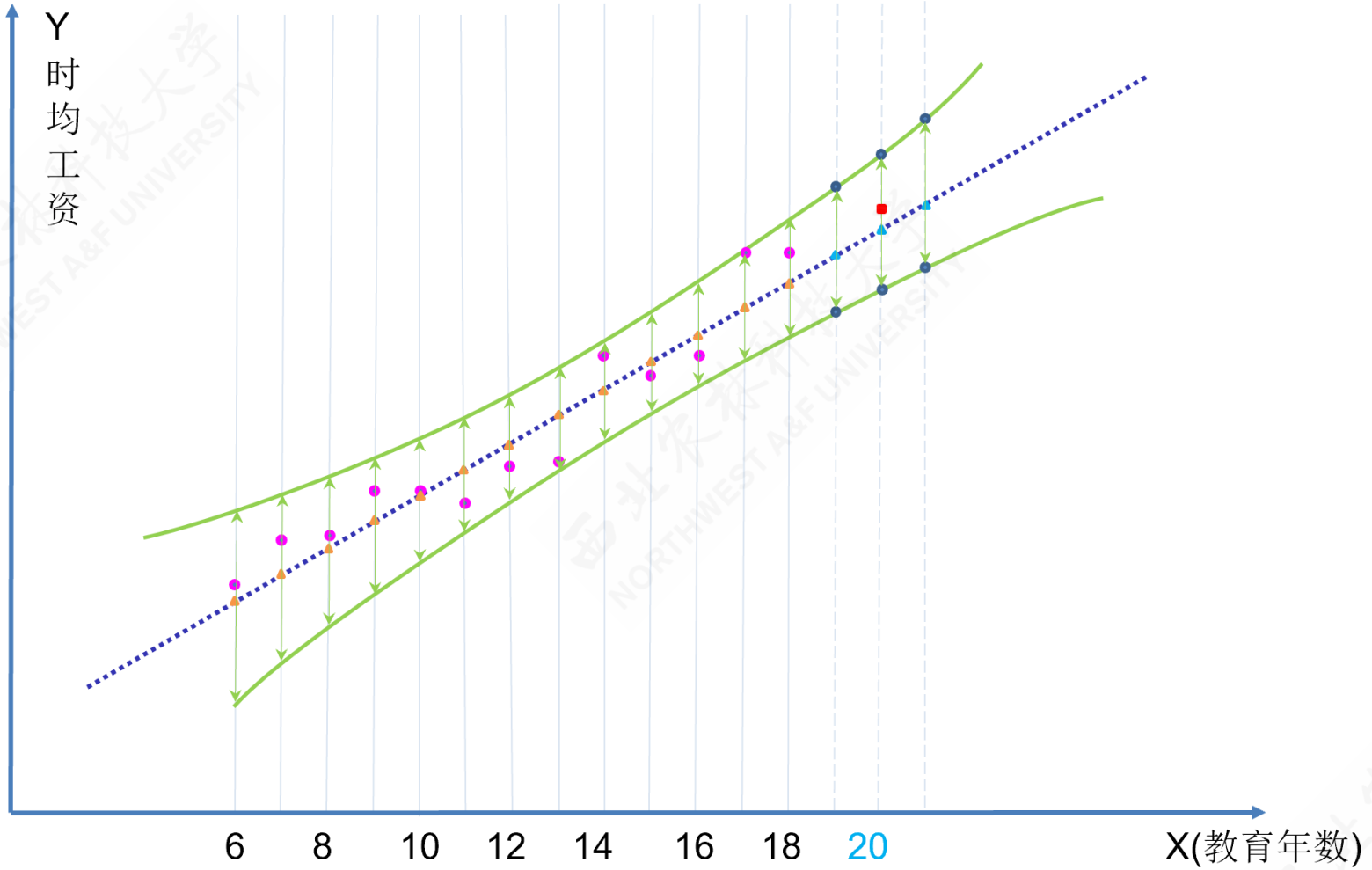
$$\hat{Y}_0 - t_{\alpha/2}(n-1)\cdot S_{(Y_0-\hat{Y}_0)} \le (Y_i|X_0) \le \hat{Y}_0 - t_{\alpha/2}(n-1)\cdot S_{(Y_0-\hat{Y}_0)}$$

置信带

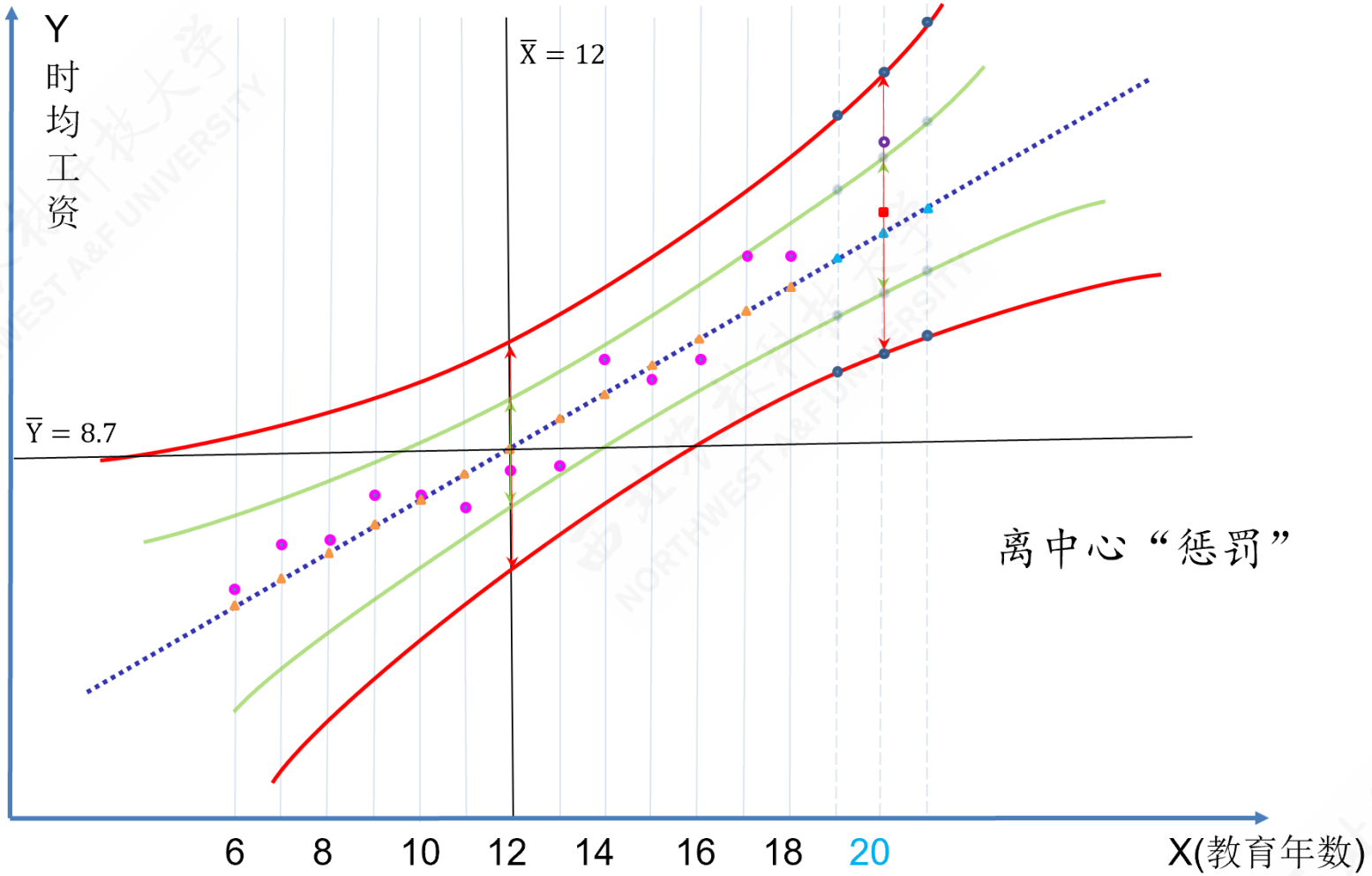
置信带(confidence interval)：对所有的X值，分别进行**均值**和**个值**分别进行预测，就能得到：

- 均值预测的置信带——总体回归函数的置信带
- 个值预测的置信带
- 预测如何可信？
  - 均值预测置信区间
  - 均值预测置信带
- 样本内置信带。——检验可靠性
- 样本外置信带。——预测未来值范围

置信带



置信带


Y
时
均
工
资
$\overline{X} = 12$
$\overline{Y} = 8.7$
离中心“惩罚”
6
8
10
12
14
16
18
20
X(教育年数)

置信带

如何理解置信带？

- 谁更宽？——均值预测更准确
- 何处最窄？—— 中心点 $(\bar{X},\bar{Y})=(12,8.67)$是历史信息的集中代表。

回归预测：总结与思考

内容总结：

- 回归预测基于一套坚实严密的“底座”：OLS估计方法、CLRM假设、BLUE估计性质
- 均值预测置信带和个值预测置信带，是对预测可信度的形象表达。
- （同等条件下）均值预测比个值预测更准确（置信带宽窄）

课堂思考：

- 同样是95%置信度区间，两个人的认识是一样的么？

课后作业：工资与教育案例扩展

- 请计算置信度 $100(1-\alpha)=95\%$下， $X_0=20$时均值的置信区间。 与 $100(1-\alpha)=90\%$时相比，有什么差异？
- 99%更值得可信么？

5.7 回归报告解读

方程表达式

表格表达式

统计软件

回归分析的形式

课程要求：会熟练、正确阅读统计软件给出的各类分析报告，理解其中的关键信息和内涵。这些分析报告包括：传统的多元回归分析报告；以及各种计量检验的辅助分析报告（如异方差white检验报告）等。

根据统计软件的不同（`stata`；`Eview`；`R`；`Excel` ……），各种分析报告呈现形式略有差异，但基本要素和信息都大抵一致。

给定如下一元回归模型：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

回归分析的形式（多行方程表达法）

形式1：多行方程表达法（整理好的**精炼报告**）：根据统计软件的原始报告，往往是选取最关键的信息，经过整理并以多行**样本回归方程**（SRF）的形式呈现，**精炼报告**的形式一般为：

$$
\begin{array}{lll}
\widehat{Y} = & -0.01 & +0.72X \\
(\mathrm{t}) & (-0.0165) & (10.4065) \\
(\mathrm{se}) & (0.8746) & (0.0696) \\
(\mathrm{fitness}) & R^2 = 0.9078; \bar{R}^2 = 0.8994 & \\
 & F^* = 108.29; p = 0.0000 &
\end{array}
$$

- 第1行表示样本回归函数（回归系数）
- 第2行(t)表示回归系数对应的**样本t统计量**（$t^*_{\hat{\beta}_i}, i \in 1, 2, \cdots, k$）
- 第3行(se)表示回归系数对应的**样本标准误差**（$S_{\hat{\beta}_i}, i \in 1, 2, \cdots, k$）
- 第4行(fitness)表示回归模型**拟合情况**和**统计检验**的简要信息，其中 R^2表示**判定系数**，$\bar{R}^2$表示**调整判定系数**，F表示模型整体显著性检验中的**样本F统计量值**（F^*），p

回归分析的形式（表格列示法）

形式2：表格列示法（整理好的**精炼报告**）：根据统计软件的原始报告，往往是选取最关键的信息，经过整理以表格形式呈现，**表格列示法**的形式呈现为：

term	estimate	std.error	statistic	p.value
(Intercept)	-0.01	0.87	-0.02	0.99
X	0.72	0.07	10.41	0.00

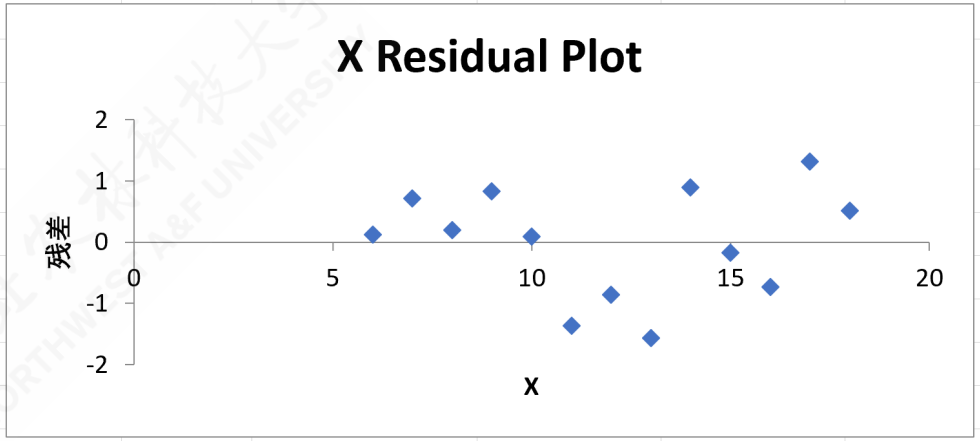
- **第1列**：`term`表示回归模型中包含的变量，也即 $X_{2i}, X_{3i}, \cdots, X_{ki}$，其中**截距项**默认为`(Intercept)`。
- **第2列**：`estimate`表示回归系数的估计值，也即 $\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k$。
- **第3列**：`std.error`表示回归系数对应的**样本标准误差**，也即 $S_{\hat{\beta}_i}, i \in 1, 2, \cdots, k$。
- **第4列**：`statistic`表示回归系数对应的**样本t统计量**，也即 $t^*_{\hat{\beta}_i}, i \in 1, 2, \cdots, k$
- **第5列**：`p.value`表示回归系数**样本t统计量**对应的概率值，也即 $Pr(t = t^*_{\hat{\beta}_i}) = p$

（示例）Excel软件原始报告：全貌

形式3：原始报告：分析软件如EViews、R、STATA、Excel等直接自动生成的多元回归分析报告。Excel软件原始分析报告形式如下：

B20

	B	C	D	E	F	G	H	I	J	K	L	M
1	X	Y										
2	6	4.4567		SUMMARY OUTPUT								
3	7	5.77										
4	8	5.9787		回归统计								
5	9	7.3317		Multiple R	0.9528							
6	10	7.3182		R Square	0.9078							
7	11	6.5844		Adjusted R Square	0.8994							
8	12	7.8182		标准误差	0.9387							
9	13	7.8351		观测值	13							
10	14	11.0223										
11	15	10.6738		方差分析								
12	16	10.8361			df	SS	MS	F	Significance F			
13	17	13.615		回归分析	1	95.43	95.43	108.29	0.00			
14	18	13.531		残差	11	9.69	0.88					
15				总计	12	105.12						
16												
17					Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
18				Intercept	-0.01	0.8746	-0.02	0.9871	-1.94	1.91	-1.94	1.91
19				X	0.72	0.0696	10.41	0.0000	0.57	0.88	0.57	0.88



（示例）Excel软件原始报告：参数估计

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	-0.01	0.8746	-0.02	0.9871	-1.94	1.91	-1.94	1.91
X	0.72	0.0696	10.41	0.0000	0.57	0.88	0.57	0.88

（示例）Excel软件原始报告：拟合优度

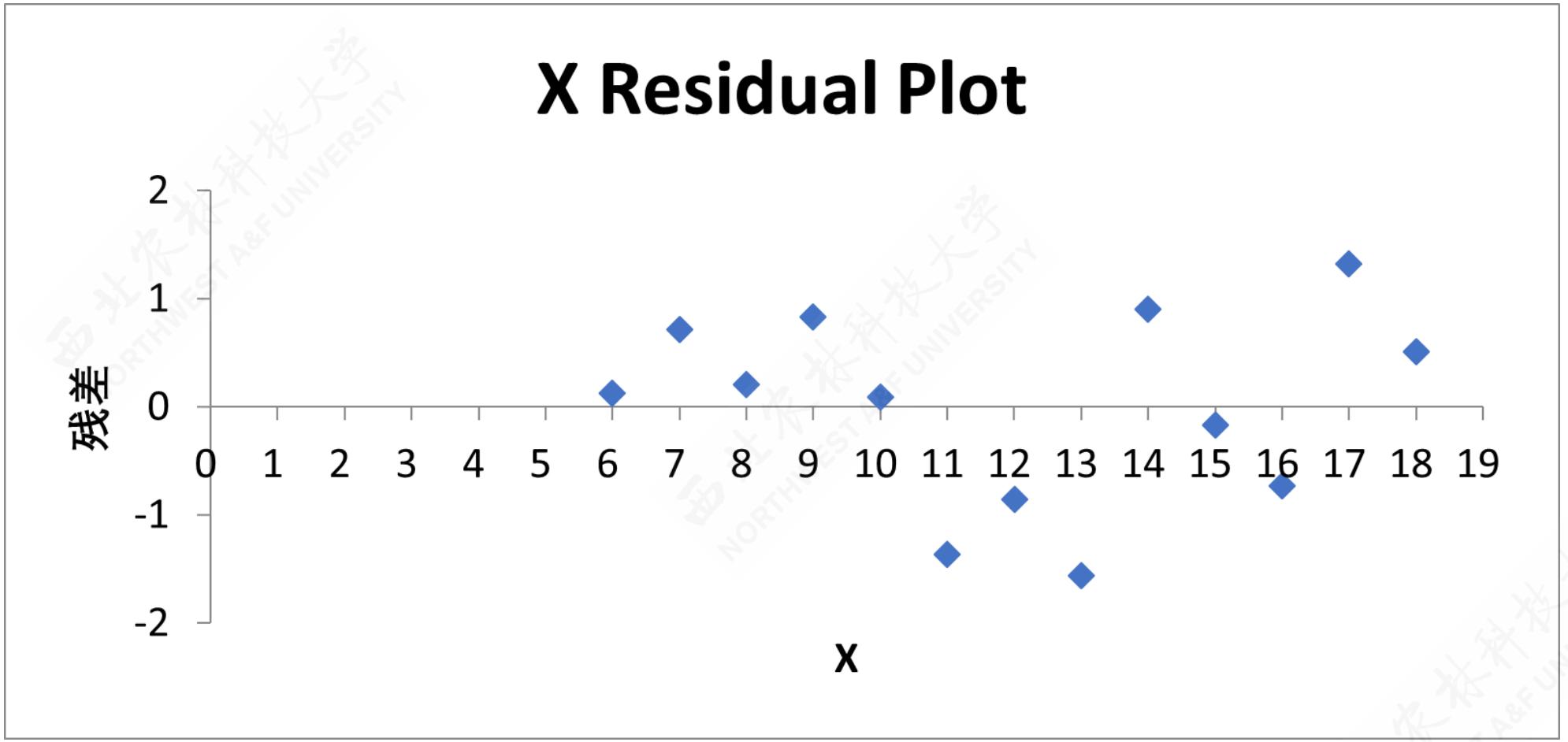
回归统计	
Multiple R	0.9528
R Square	0.9078
Adjusted R Square	0.8994
标准误差	0.9387
观测值	13

（示例）Excel软件原始报告：方差分解

方差分析					
	df	SS	MS	F	Significance F
回归分析	1	95.43	95.43	108.29	0.00
残差	11	9.69	0.88		
总计	12	105.12			

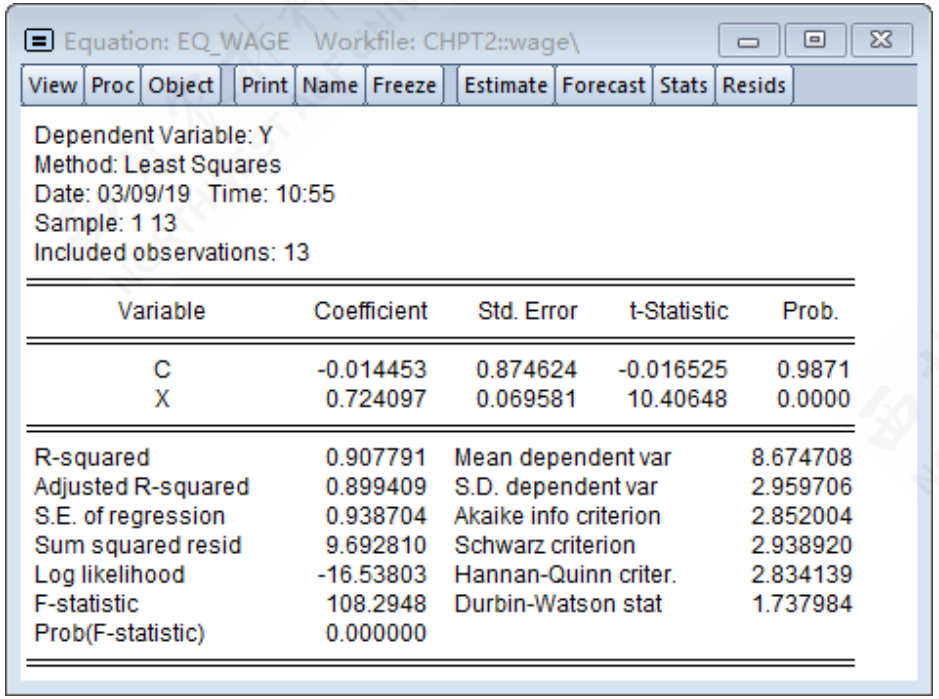
（示例）Excel软件原始报告：残差表

（示例）Excel软件原始报告：残差图



回归分析的形式（EViews软件原始报告）

形式3：原始报告：分析软件如EViews、R、STATA等直接自动生成的多元回归分析报告。EViews软件原始分析报告形式如下：**抬头区域**



- `Dependent Variable: Y`：因变量
- `Method: Least Squares`：分析方法
- `Date: 03/09/19 Time: 10:55`：分析的时间
- `Sample: 1 13`：样本范围
- `Included observations: 13`：样本数n

回归分析的形式（EViews软件原始报告）

形式3：原始报告：分析软件如EViews、R、STATA等直接自动生成的多元回归分析报告。EViews软件原始分析报告形式如下：**三线表区域**

Equation: EQ_WAGE Workfile: CHPT2::wage\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: Y
Method: Least Squares
Date: 03/09/19 Time: 10:55
Sample: 1 13
Included observations: 13

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.014453	0.874624	-0.016525	0.9871
X	0.724097	0.069581	10.40648	0.0000

R-squared	0.907791	Mean dependent var	8.674708
Adjusted R-squared	0.899409	S.D. dependent var	2.959706
S.E. of regression	0.938704	Akaike info criterion	2.852004
Sum squared resid	9.692810	Schwarz criterion	2.938920
Log likelihood	-16.53803	Hannan-Quinn criter.	2.834139
F-statistic	108.2948	Durbin-Watson stat	1.737984
Prob(F-statistic)	0.000000		

- **第1列**：Variable表示模型包含的变量，$X_{2i}, X_{3i}, \cdots, X_{ki}$，其中**截距项**默认为C。
- **第2列**：Coefficient回归系数，也即$\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k$；
- **第3列**：Std. Error回归系数的样本标准误差，也即也即$S_{\hat{\beta}_i}, i \in 1, 2, \cdots, k$。
- **第4列**：t-Statistic表示回归系数对应的**样本t统计量**，也即$t^*_{\hat{\beta}_i}, i \in 1, 2, \cdots, k$；

回归分析的形式（EViews软件原始报告）

形式3：原始报告：分析软件如EViews、R、STATA等直接自动生成的多元回归分析报告。EViews软件原始分析报告形式如下：**指标值区域（左）**

Equation: EQ_WAGE Workfile: CHPT2::wage\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: Y
Method: Least Squares
Date: 03/09/19 Time: 10:55
Sample: 1 13
Included observations: 13

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.014453	0.874624	-0.016525	0.9871
X	0.724097	0.069581	10.40648	0.0000

R-squared	0.907791	Mean dependent var	8.674708
Adjusted R-squared	0.899409	S.D. dependent var	2.959706
S.E. of regression	0.938704	Akaike info criterion	2.852004
Sum squared resid	9.692810	Schwarz criterion	2.938920
Log likelihood	-16.53803	Hannan-Quinn criter.	2.834139
F-statistic	108.2948	Durbin-Watson stat	1.737984
Prob(F-statistic)	0.000000		

- R-squared：回归模型**判定系数** R^2。
- Adjusted R-squared：回归模型**调整判定系数** $\bar{R}^2$。
- S.E. of regression：回归模型的**回归误差标准差** $\hat{\sigma}$。
- Sum squared resid：回归模型的**残差平方和RSS** $RSS=\sum e_i^2$。
- Log likelihood：回归模型的**对数似然值**。

回归分析的形式（EViews软件原始报告）

形式3：原始报告：分析软件如EViews、R、STATA等直接自动生成的多元回归分析报告。EViews软件原始分析报告形式如下：**指标值区域（右）**

Equation: EQ_WAGE Workfile: CHPT2::wage\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: Y
Method: Least Squares
Date: 03/09/19 Time: 10:55
Sample: 1 13
Included observations: 13

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.014453	0.874624	-0.016525	0.9871
X	0.724097	0.069581	10.40648	0.0000

R-squared	0.907791	Mean dependent var	8.674708
Adjusted R-squared	0.899409	S.D. dependent var	2.959706
S.E. of regression	0.938704	Akaike info criterion	2.852004
Sum squared resid	9.692810	Schwarz criterion	2.938920
Log likelihood	-16.53803	Hannan-Quinn criter.	2.834139
F-statistic	108.2948	Durbin-Watson stat	1.737984
Prob(F-statistic)	0.000000		

- Mean dependent var：Y的**均值** $\bar{Y}$。
- S.D. dependent var：Y的**样本标准差** S_Y。
- Akaike info criterion：回归模型的**AIC信息准则**。
- Schwarz criterion：回归模型的**Schwarz准则**。
- Hannan-Quinn criter.：回归模型的**Hannan-Quinn准则**。

回归分析的形式（R软件原始报告）

形式4：原始报告：分析软件如EViews、R、STATA等直接自动生成的多元回归分析报告。R软件原始分析报告形式如下：

```
Call:
lm(formula = mod_wage, data = data_wage)

Residuals:
   Min     1Q Median     3Q    Max
-1.564 -0.735  0.127  0.716  1.320

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  -0.0145     0.8746   -0.02      0.99
X             0.7241     0.0696   10.41 0.0000005 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.94 on 11 degrees of freedom
Multiple R-squared:  0.908,    Adjusted R-squared:  0.899
F-statistic:  108 on 1 and 11 DF,  p-value: 0.000000496
```

本章结束