本章内容：

第二章 一元线性回归模型

- 2.1 回归分析的性质
- 2.2 一元线性回归：一些基本思想
- 2.3 一元线性回归模型：参数估计
- 2.4 一元线性回归模型：经典正态线性假设
- 2.5 一元线性回归模型：区间估计和假设检验
- 2.6 一元线性回归模型：一些扩展

本节内容：

§2.1 回归分析的性质

§ 2.1.1“回归”的历史渊源

高尔顿的发现：

- **普遍回归定律（Law of universal regression）：**
  - 父母高，子女也高；父母矮，子女也矮；
  - 给定父母的身高，子女的**平均身高**趋向于（回归）全体人口的**平均身高**。

皮尔逊的证实：

- **数据：部分家庭群体的一千多名成员的身高记录。**
- **研究分组：父亲高的群体VS父亲矮的群体**
  - 父亲高的群体，子辈平均身高要低于其父辈；
  - 父亲矮的群体，子辈平均身高要高于其父辈。

- **高尔顿对回归的解释：**

  “回归到中等(regression to mediocrity)”

§ 2.1.2 “回归”的现代解释

定义2.1.2:

- **关键问题：给定父辈身高，子辈身高如何变化。**
  - 给定变量（解释变量，X）：父辈身高
  - 目标变量（被解释变量，Y）：子辈身高
  - 变量之间的关系：X→Y
  - 问题：关系能否够反过来？比如 Y→X

- **回归分析的本质：**

回归分析是关于研究被解释变量（Y）对另一个或多个解释变量(X)的依赖关系，其用意在于：通过解释变量X（在重复抽样中）的已知或设定值，去估计和（或）预测被解释变量Y（总体）均值。

案例说明：

Case 1:
- 给定父亲身高，在一个假想人口总体中的子辈身高分布。

Case 2:
- 给定年龄，男孩子身高总体的分布。

Case 3:
- 给定税后或可支配收入，个人消费是如何分布的；这种分析有助于估计边际消费倾向（MPC），就是实际收入每美元价值的变化所引起的消费支出的平均变化。

Case 4:
- 失业率是怎样影响货币工资变化的。

Case 5:
- 通货膨胀率如何影响人们以货币形式持有的收入比例的变化。

- **共同之处：** *给定X变量（解释变量），Y变量（被解释变量）如何变化。*

Case 1:
给定父亲身高，子辈在一个假想人口总体中的身高分布。

Figure 1.1

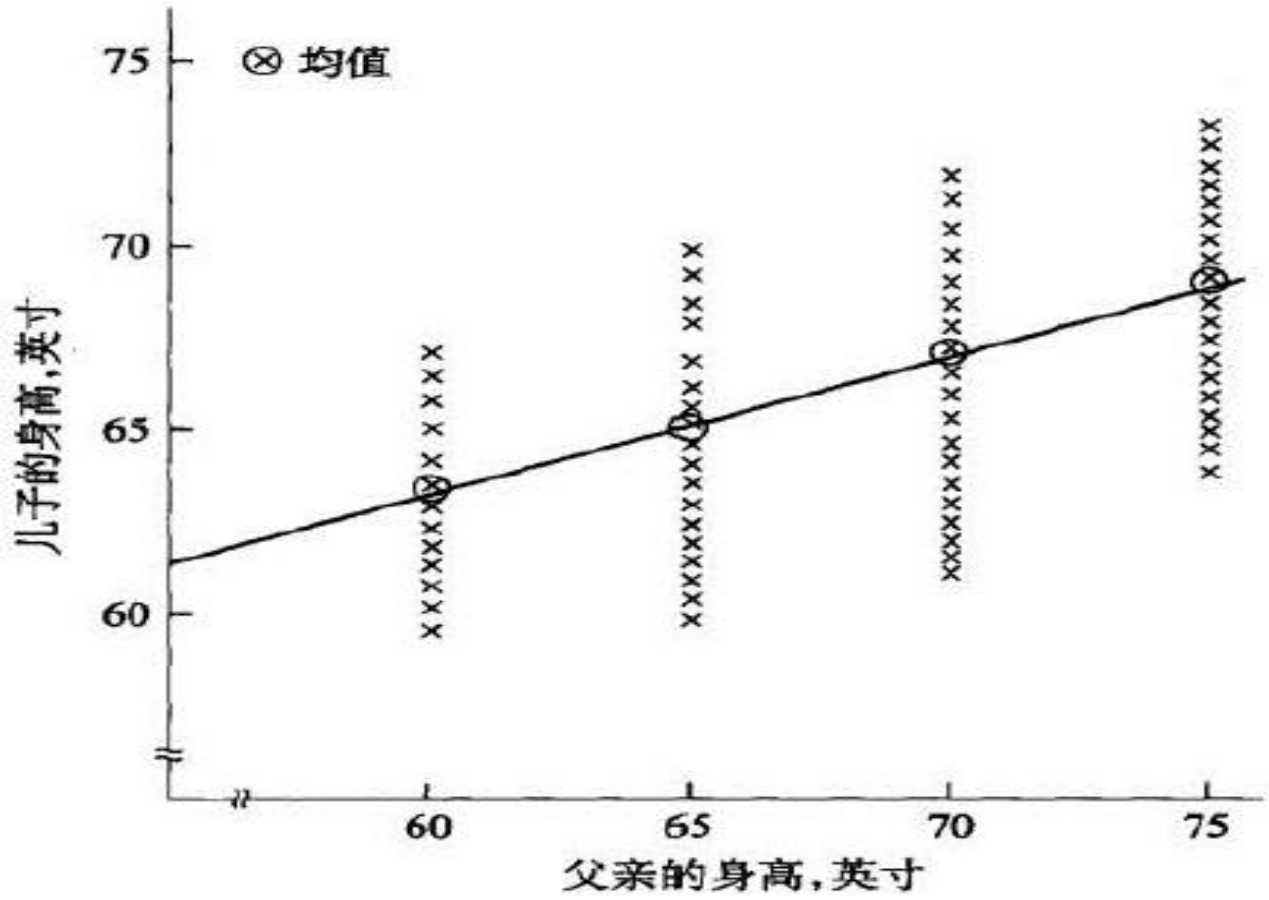


图1-1 给定父亲身高时儿子身高的假想分布

Case 2:
给定年龄，男孩子身高总体的分布。

Figure 1.2

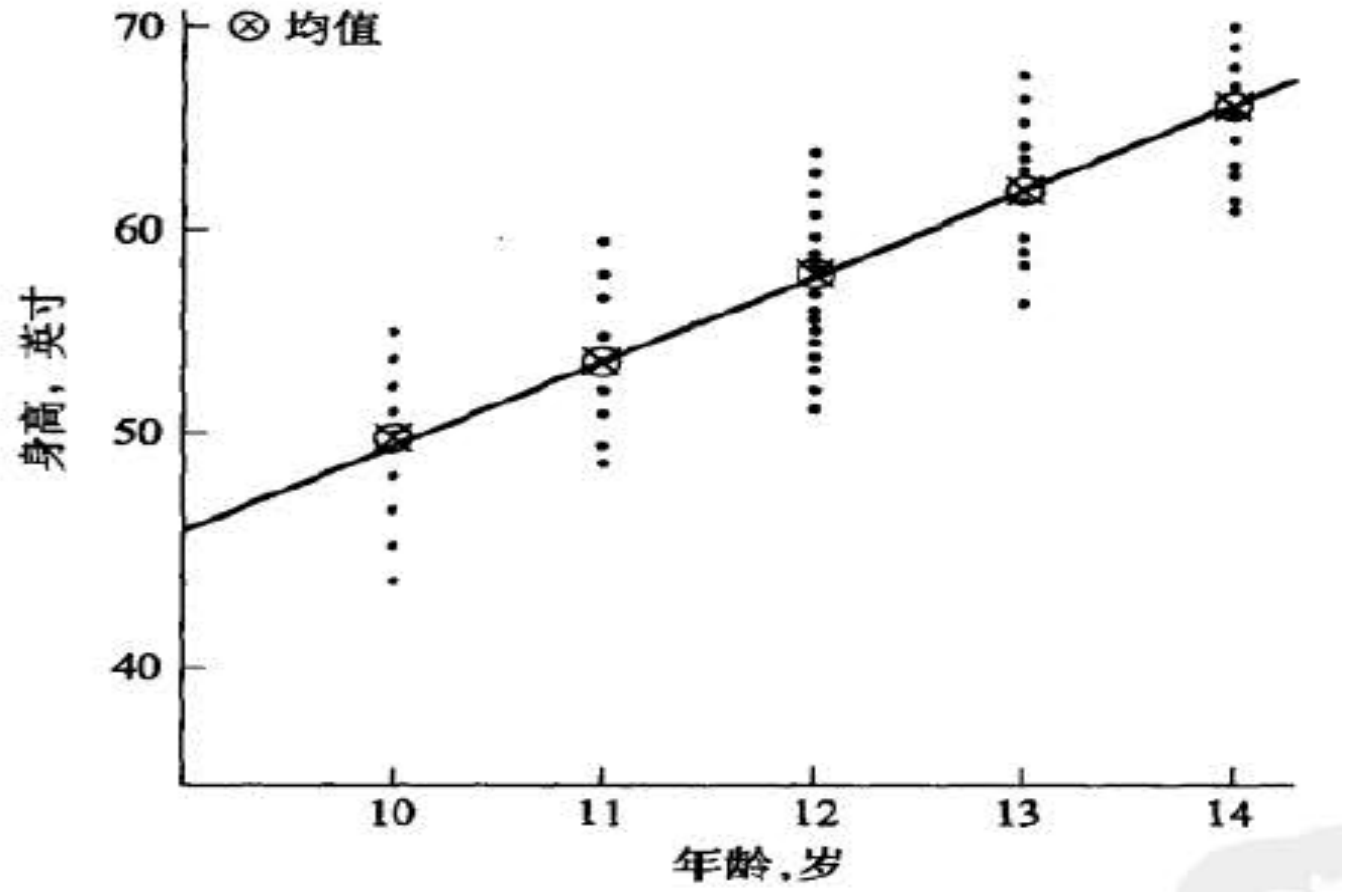


图1-2 对应于选定年龄的假想身高分布

Case 4:
失业率是怎样影响货币工资变化的

Figure 1.3

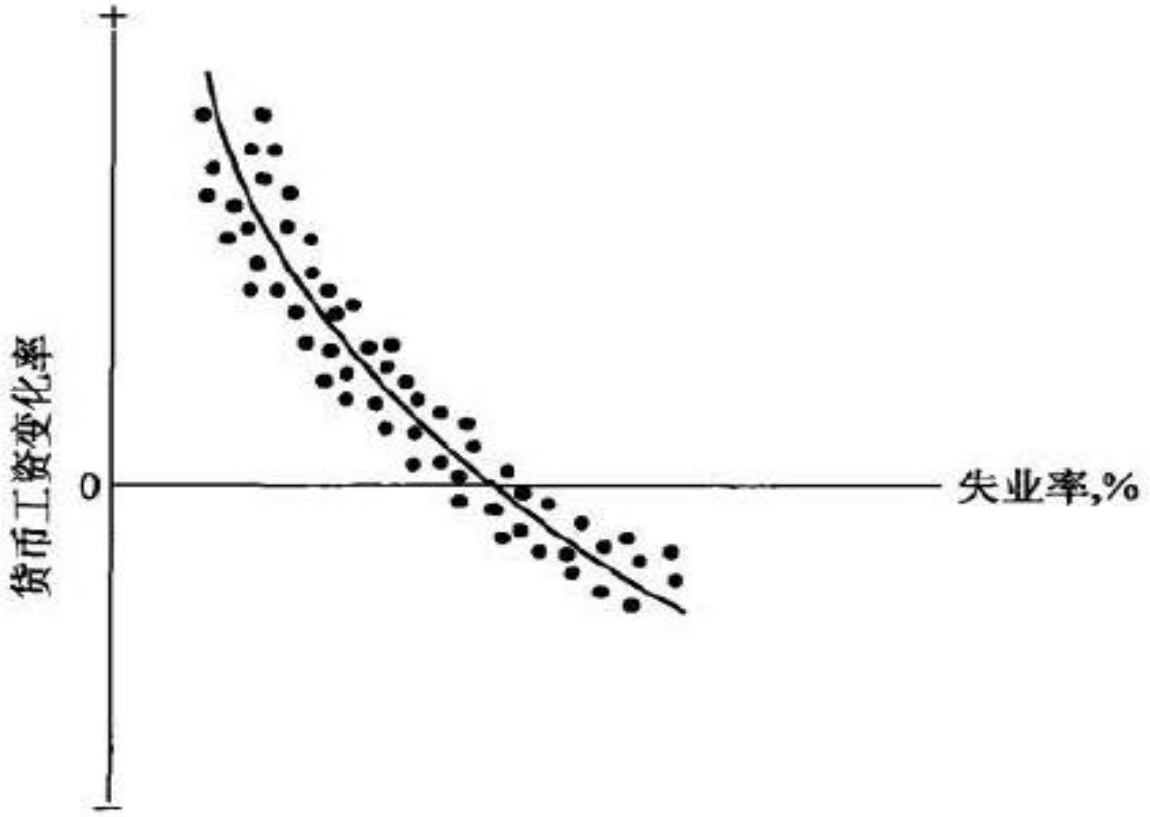


图1-3 假想的菲利普斯曲线

Case 5:
通货膨胀率如何影响人们以货币形式持有的收入比例变化

Figure 1.4

根据货币经济学，其他条件不变，通货膨胀率π 越高，人们愿意以货币形式持有的收入比例k越低

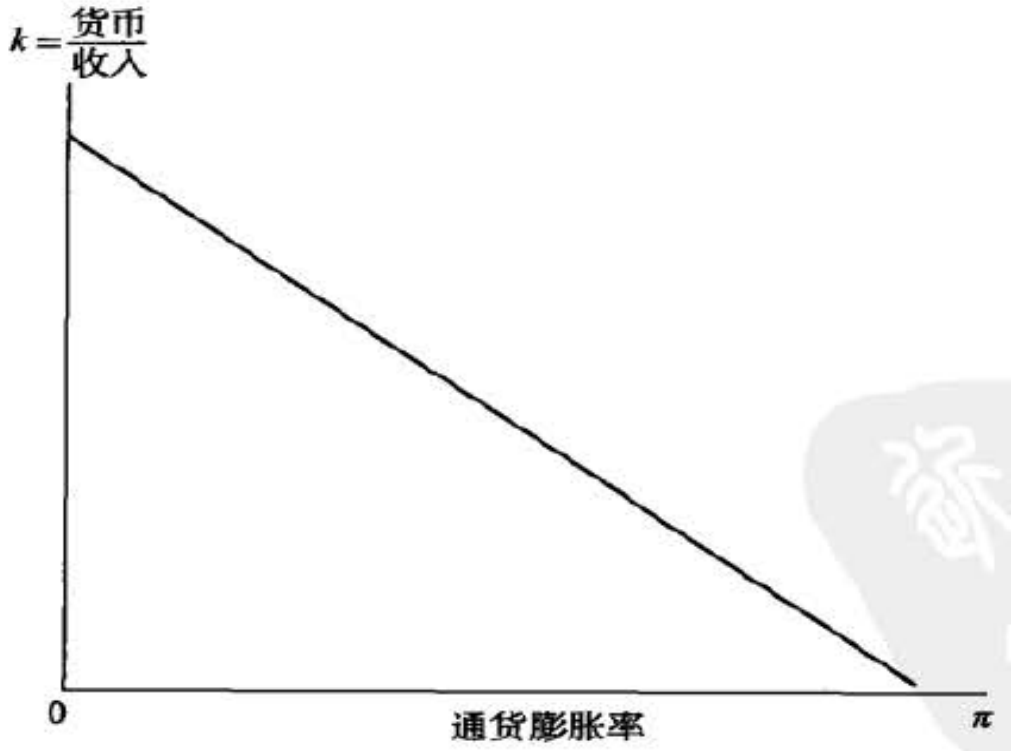


图1-4 货币持有与通货膨胀率 π 之间的关系

§ 2.1.3 确定性关系与不确定性关系

思考：

- **关系的起源：**
  - 哲学起点：普遍联系（哲学论断）
  - 确定性关系是相对的，不确定关系是绝对的；
  - 人类认知规律：从不确定性关系开始探索，直到发现确定性关系。

- **不确定性关系：**
  - 不确定性关系：变量之间的随机性关系。
  - 统计关系：经过统计方法验证的不确定性关系。
  - 例1:某产品的需求量与收入、产品价格、替代品价格之间的关系。
  - 例2:农作物产量与气温、降雨量、阳光及施肥量之间的关系。

- **确定性关系：**
  - 确定性关系：变量之间经过验证的稳定关系，可表示为数学上的函数关系。如，经典物理学中的关系：
    - 万有引力关系：$F = km_1 \cdot m_2 / r^2$
    - 欧姆定律：$I = V/k$

§ 2.1.5 从相关关系到回归关系

- **相关关系：**
  - 统计学分析：测度两个变量X, Y之间的**线性**关联程度（双向的相关关系）。
  - 方法：计算变量之间的相关系数（correlation coefficient)
  - 实质：变量之间双向的、不确定的统计关系。
  - 特别说明：相关系数为0仅表明不存在线性相关，不能否认非线性相关的情形。例如变量之间存在U型关系。

- **回归关系：**
  - 回归（计量）分析：被解释变量Y与解释变量X之间的单向的相关关系。（X→Y）
  - 方法：建立回归模型，分析给定X变量（解释变量），Y变量（被解释变量）如何变化。
  - 实质：变量之间单向的、不确定的统计关系。

§ 2.1.5 从相关关系到回归关系

- **相关关系与回归关系比较：**
  - **共同点：**都是变量间的不确定的统计关系。
  - **区别1：**
    - 相关关系侧重两个随机变量之间的**关系强度**；
    - 回归关系侧重两个或多个变量之间的**具体关系，**即给定X, Y的变化如何受X影响。
  - **区别2：**
    - 相关关系是双向的，具有**对称性**；
    - 回归关系是单向的，具有**非对称性**。

§ 2.1.4 从回归关系到因果关系

- **回归关系：**
  - 实质：被解释变量Y与解释变量X之间单向的、不确定的统计关系。
  - 证明方式：回归模型和统计检验。
- **因果关系（广义）：**
  - 实质：变量之间的内在的先后逻辑关系。
  - 验证方法：经验判断和理性思考。
  - 分类（依据是否确定）：
    - 因果关系（狭义，通常）：反复验证过的确定关系；
    - 因果机制（科学分析）：部分验证的不确定关系。

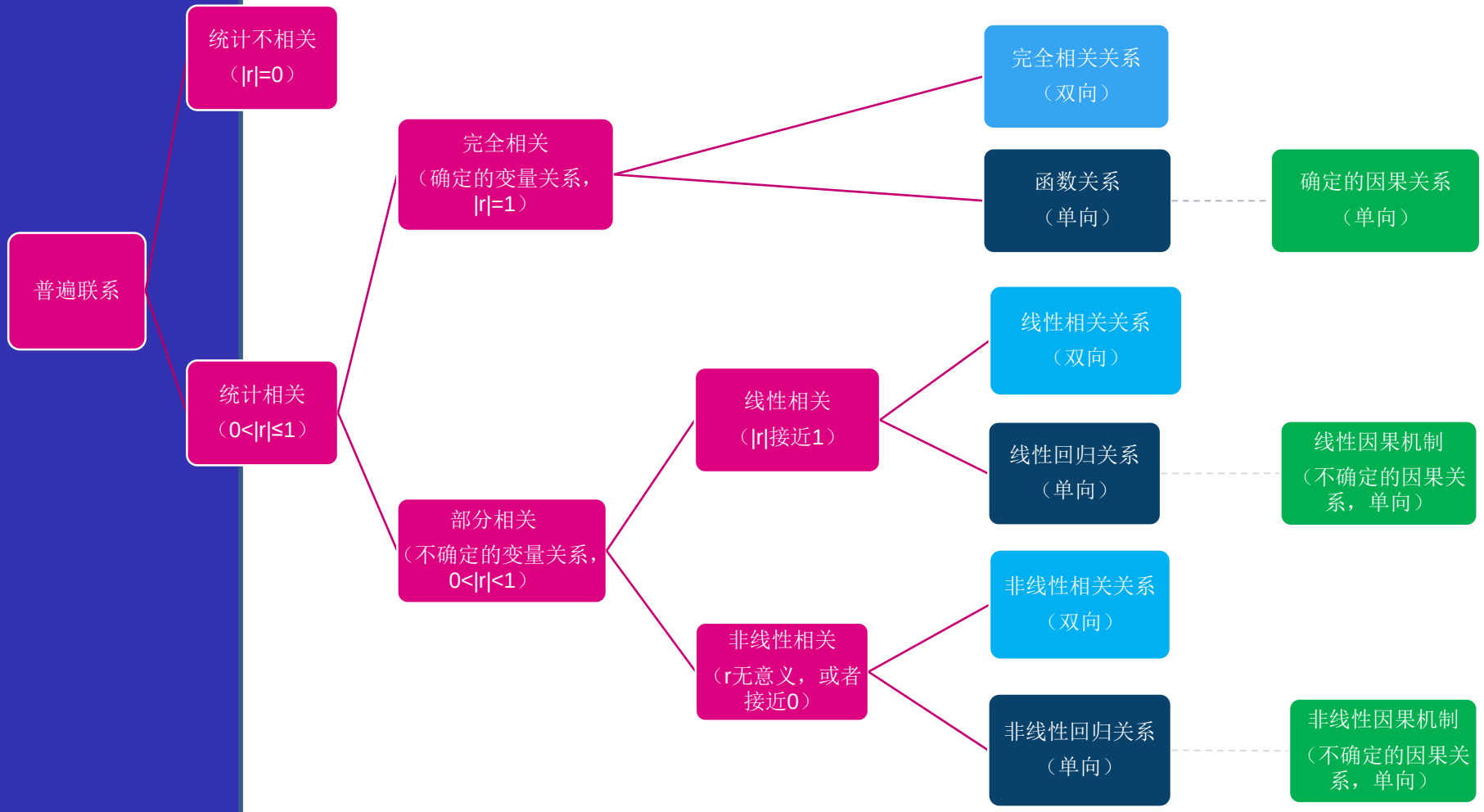
§ 2.1.4 从回归关系到因果关系

- **回归关系与因果关系的比较：**
  - 区别：
    - 回归关系是（外在）可观测变量之间的统计关系，可以通过统计和计量方法检验和验证；
    - 因果关系是内在逻辑关系，难以直接观测，只能通过经验判断和理性思考来把握；
  - 关联：
    - 回归关系的存在并不一定意味着因果关系；
    - 若因果关系存在，则必定表现为某种显著的回归关系。

“一个统计关系式，不管多强，也不管多么有启发性，永远不能确定因果方面的联系：对因果关系的理念，必须来自统计学意外，最终来自这种或那种理论。”

——Kendall & Stuart

总结：
从相关到因果


普遍联系
统计不相关
（|r|=0）
统计相关
（0<|r|≤1）
完全相关
（确定的变量关系，
|r|=1）
部分相关
（不确定的变量关系，
0<|r|<1）
线性相关
（|r|接近1）
非线性相关
（r无意义，或者
接近0）
完全相关关系
（双向）
函数关系
（单向）
确定的因果关系
（单向）
线性相关关系
（双向）
线性回归关系
（单向）
线性因果机制
（不确定的因果关系，单向）
非线性相关关系
（双向）
非线性回归关系
（单向）
非线性因果机制
（不确定的因果关系，单向）

§ 2.1.6 术语与符号

Table 2.1.2
（术语表）

Y	X
因变量(Dependent variable)	**自变量（Independent variable）**
被解释变量(Explained variable)	**解释变量(Explanatory variable)**
预测子(Predictand)	预测元(Predictor)
回归子(Regressand)	回归元(Regressor)
响应变量(Response variable)	刺激变量 (Stimulus variable)
内生(Endogenous)变量	**外生(Exogenous)变量**
结果(Outcome) 变量	协变量(Covariate)
被控变量(Controlled variable)	控制变量(Control variable)

定义1：

- 双变量回归分析(two-variable regression analysis)：
  - 研究一个变量对仅仅一个解释变量的依赖关系，如消费支出对实际收入的依赖关系

定义2：

- 多元回归分析(multiple regression analysis)：
  - 在研究一个变量对多于一个解释变量的依赖关系，如农作物收成依赖于气温、降雨量、阳光和施肥量等；

定义3：

- 随机
  - “Random” 和“Stochastic” 这两个单词是同义语，都是随机的意思。
  - 一个随机变量是指这样的一个变量:它以给定的概率取任一特定数值，可正可负。

约定1：
（模型符号）

- 因变量：Y
- 自变量：$X(X_1, X_2, \dots, X_k)$，其中：
  - $X_k(X_{k1}, X_{k2}, \dots, X_{ki})$代表第k个解释变量，下标i（或t）则指第i（或t）次观测值。
    - 横截面数据(cross-sectional data)：用观测值下标i来表示，这是指在一个时间点上搜集的数据。
    - 时间序列数据(time series data)，用下标t来表示，这是一个时期内收集的数据。

约定2：
（情境符号）

- 观测值数量：
  - 总体容量：即总体中的观测值总个数
    - N（横截面数据下使用）
    - T（时间序列数据下使用）
  - 样本容量：即样本中的观测值总个数
    - n（横截面数据下使用）
    - t（时间序列数据下使用）

§ 2.1.7 数据的性质与来源

Type1：时间序列数据(time series data)：

定义1：

- **时间序列：对一个变量在不同时间取值的一组观测结果。**
  - 实时牌价：如股票价格
  - 每日(daily)：如天气预报
  - 每周(weekly)：如货币供给数字
  - 每月(monthly)：如失业率和消费者价格指数
  - 每季度(quarterly)：如GDP，
  - 每年(annually)：如政府预算
  - 每5年(quinquennially)：如制造业普查资料
  - 每10年(decennially)：如人口普查资料

定义2：

- **平稳性(stationary)问题。**
  - 如果一个时间序列的均值和方差不随时间而系统地变化，那它就是平稳的(stationary) 。

Type1:
时间序列数据(time series data):

Figure 1-5:

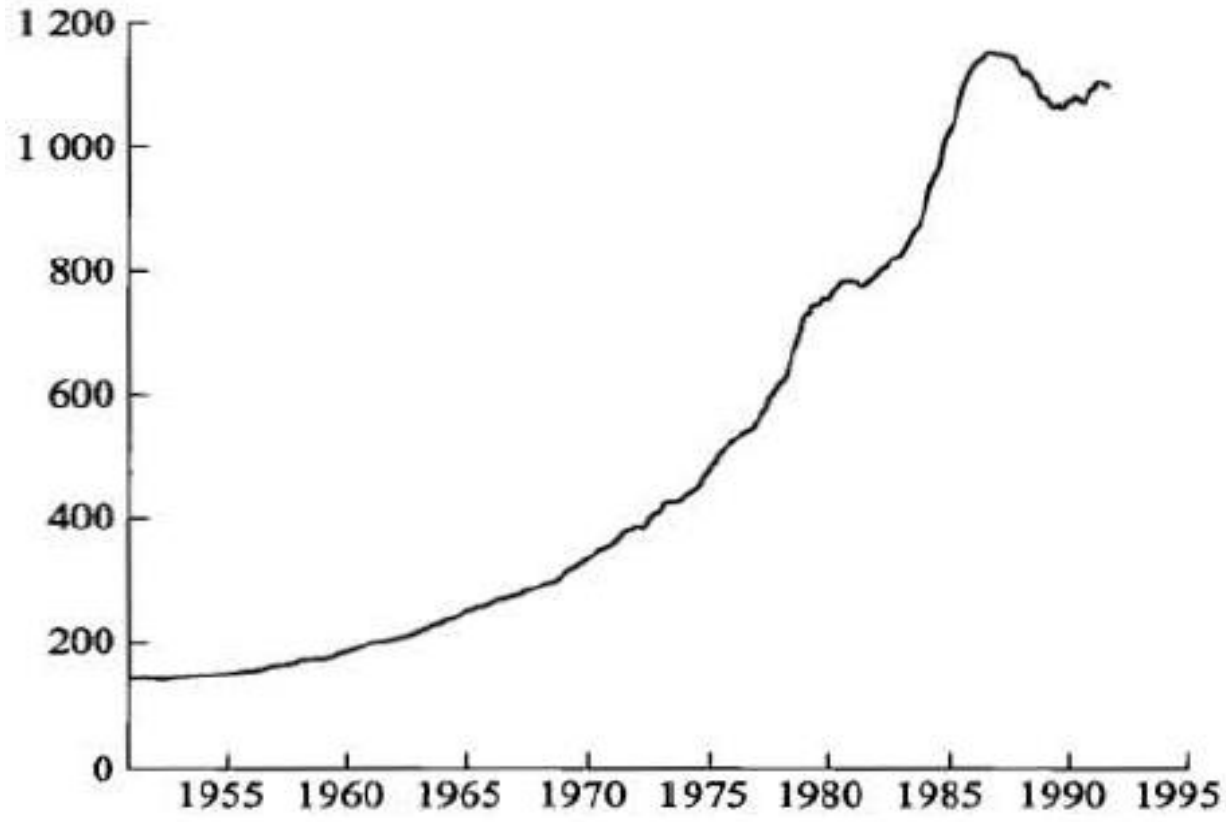


图1-5：1951 年1 月-1999 年9 月美国的M1货币供给

Type2:
截面数据(cross-section data):

定义1:

- 横截面数据：对一个或多个变量在同一时间点上收集的数据

定义2:

- 异质性(heterogeneity)：
  - 当我们的统计分析包含有异质的单位时，我们必须考虑尺度(size)或规模效应(scale effect)以避免造成混乱。

Type2:
截面数据(cross-section data):

Table1-1:

表1-1 美国蛋类生产

state	Y1	X1	Y2	X2
AL	2206	92.7	2186	91.4
AK	0.7	151	0.7	149
AR	3620	86.3	3737	91.8
CA	7472	63.4	7444	58.4
...	...	...	...	...
VA	934	86.3	988	81.2
WA	1287	74.1	1313	71.5
WI	910	60.1	873	54
WY	1.7	83	1.7	83

注:
State:美国50个洲的缩写
Yl = 1990 年鸡蛋产量(百万个);
X1 = 1990 年每打鸡蛋的价格(美分)
Y2 =1991 年鸡蛋产量(百万个)
X2 =1991 年每打鸡蛋的价格(美分)

Figure 1-6:

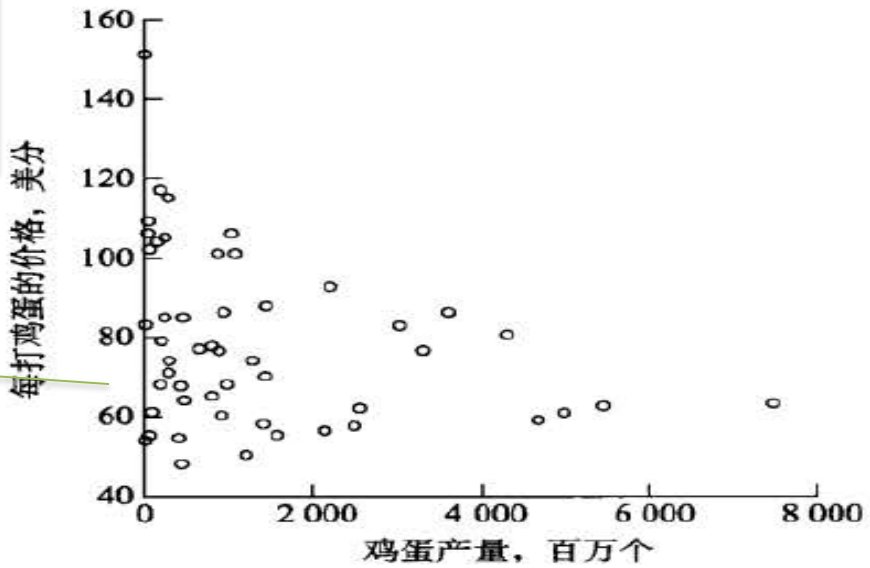


图1-6 1990 年蛋产量与价格的关系

思考:

图中体现了怎样的规模效应?

Type3：
面板数据(Panel Data)：

定义1：

定义2：

定义3：

- 面板数据：是兼有时间序列和横截面数据两种成份，指对相同的横截面单元在时间轴上进行跟踪调查的数据。
  - 平衡面板(balanced panel)：所有截面单元都具有相同的观测次数
  - 非平衡面板(unbalanced panel)：并非所有截面单元都具有相同的观测次数
- 数据点（观测数）：截面单元数*时期数
- “平稳性”问题：
- “异方差”问题：

Type3:
面板数据(Panel Data):

GE:通用；US:美国钢铁
真实总投资(I)
一年前的企业真实价值(F_{-1})
一年前的真实资本存量(C_{-1})

Table 1-2:

表1-2 1935--1954 年间美国两大公司的投资敏捷

GE				US			
年份	I	F_{-1}	C_{-1}	年份	I	F_{-1}	C_{-1}
1935	33.1	1170.6	97.8	**1935**	209.9	1362.4	53.8
1936	45	2015.8	104.4	**1936**	355.3	1807.1	50.5
1937	77.2	2803.3	118	**1937**	469.9	2673.3	118.1
1938	44.6	2039.7	156.2	**1938**	262.3	1801.9	260.2
1939	48.1	2256.2	172.6	**1939**	230.4	1957.3	312.7
1940	74.4	2132.2	186.6	**1940**	361.6	2202.9	254.2
1941	113	1834.1	220.9	**1941**	472.8	2380.5	261.4
1942	91.9	1588	287.8	**1942**	445.6	2168.6	298.7
1943	61.3	1749.4	319.9	**1943**	361.6	1985.1	301.8
1944	56.8	1687.2	321.3	**1944**	288.2	1813.9	279.1
1945	93.6	2007.7	319.6	**1945**	258.7	1850.2	213.8
1946	159.9	2208.3	346	**1946**	420.3	2067.7	232.6
1947	147.2	1656.7	456.4	**1947**	420.5	1796.7	264.8
1948	146.3	1604.4	543.4	**1948**	494.5	1625.8	306.9
1949	98.3	1431.8	618.3	**1949**	405.1	1667	351.1
1950	93.5	1610.5	647.4	**1950**	41		
1951	135.2	1819.4	671.3	**1951**	58		
1952	157.3	2079.7	726.1	**1952**	64		
1953	179.5	2371.6	800.3	**1953**	6		
1954	189.6	2759.9	888.9	**1954**	45		

思考：

问1：平衡面板还是非平衡面板？
问2：多少数据点？
问3：两个公司投资函数是否相同？

“研究结果不可能比数据的质量更好”

定义1：

定义2：

定义3：

- 经济分析中的数据来源：
  - 非实验性的数据：社会科学中的数据
    - 如GDP、失业、股票价格等。
  - 实验性的数据：自然科学中的数据
    - 例如研究肥胖对血压的影响时，要对饮食、烟酒习惯等变量进行控制（相同情况下）
- 经济分析中数据的准确性：
  - 非实验性数据中的观测误差、疏漏
  - 实验性数据中的测量误差，如源自近似计算或进位
  - 选择性偏误(selectivity bias)，如源自问卷调查中无应答(nonresponse）
  - 抽样方法不同，导致样本比较分析的标准不一致
  - 数据加总问题，它未必能揭示太多有关个人或微观单位的情况, 如GNP、就业、通货膨胀、失业等数据
  - 数据保密问题，某些数据只能以抽象形式公布

§ 2.1.8 观测变量的测量尺度

定义1：

定义2：

- **比率尺度(ratio scale)：**
  - **比率尺度变量**具备如下三个性质：对于一个变量X，取其两个值X1和X2，**比率**(比率属性)X1/X2 和**距离**(距离属性)(X2-X1) 都是有意义的量。此外，这些值在这种尺度下存在着一种自然**顺序**（上升或下降）(顺序属性) 。因此如X2≤X1或X2≥X1之类的比较也是有意义的。
    - 如：GDP(亿元)、个人收入(元)等
- **区间尺度(interval scale) ：**
  - **区间尺度变量**满足比率尺度变量的顺序属性和距离属性，但不满足比率属性。
    - 两个时期之内的距离(如2000 – 1995) 是有意义的，但两个时期的比率(2000/1995) 就没有什么意义。
    - 2013年8 月11 日上午11 点天气预报说杨凌的温度是华氏60 度，而长沙达到华氏90 度。说长沙比杨凌暖和50%没有意义，所以，温度不是比例尺度。这主要是因为华氏温标不是以0 度作为起点所致。

定义3：

- 序数尺度(ordinal scale)
  - 只要一个变量满足比率尺度的顺序属性（即自然顺序），那它就属于**序数尺度变量**。
    - 五分量表
    - 无差异曲线

定义4：

- 名义尺度(nominal scale)
  - **名义尺度变量**只表示不同的类别，它不具备比率尺度变量的任何一个属性。
    - 如性别(男、女）和婚姻状况(已婚、未婚、离婚、分居〉之类的变量。

测量尺度的类型、属性及变量类型对应关系

测量尺度的类型、属性及变量类型对应表

		测量尺度属性			变量类型
		顺序属性	距离属性	比率属性	
测量尺度类型	比率尺度	√	√	√	数值型变量
	区间尺度	√	√	×	
	序列尺度	√	×	×	顺序型变量
	名义尺度	×	×	×	分类型变量

注："√"表示测量尺度**具备**该属性，"×"表示测量尺度**不具备**该属性。

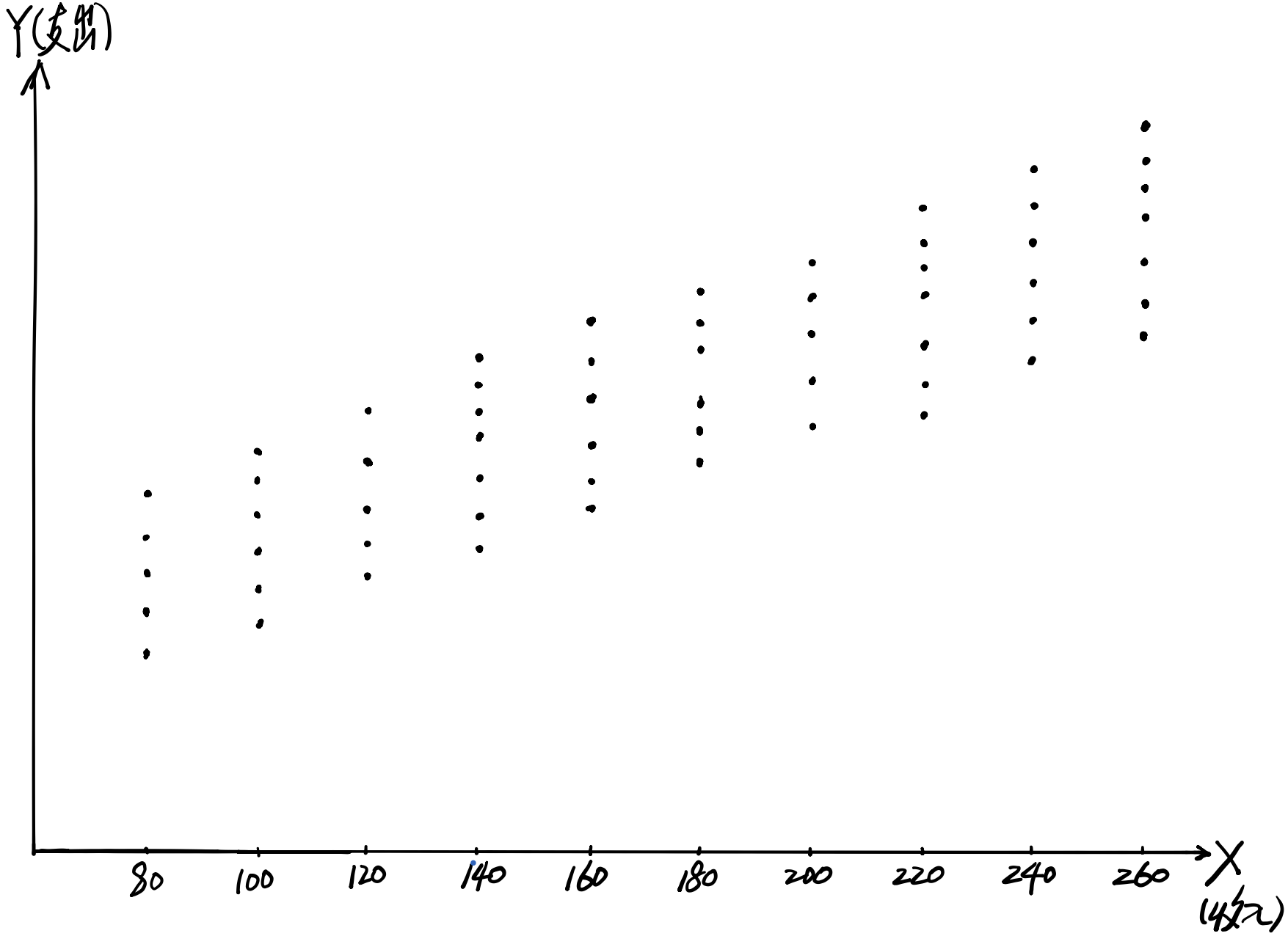
本节内容：

§2.2 双变量回归分析：一些基本思想

60个家庭的收入和支出情况——假设的总体

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**									

问1：总体是什么？有多少总体单位？


Y(支出)
80
100
120
140
160
180
200
220
240
260
X
(收入)

条件概率和条件期望

定义1：

- **无条件概率：**不受X变量取值影响下，Y_i出现的可能性。

$$p(Y_i)$$

- **条件概率：**给定变量X的取值条件下，Y_i出现的可能性。

$$p(Y_i \mid X)$$

定义2：

- **无条件期望：**不受X变量取值影响下，变量Y的期望值。

$$\mathrm{E}(Y)=\sum_{i=1}^{N} Y_i \cdot p(Y_i) \quad \text{(离散)}$$

定义3：

$$\mathrm{E}(Y)=\int Y \cdot g(Y)dY \quad \text{(连续)}$$

- **条件期望：**在给定变量X的取值条件下，Y的期望值。

$$\mathrm{E}(Y \mid X)=\sum_{i=1}^{N}(Y_i \mid X)\cdot p(Y_i \mid X) \quad \text{(离散)}$$

$$\mathrm{E}(Y_i|\mathrm{X})=\int (Y_i|\mathrm{X})\cdot g(Y_i|\mathrm{X})dY \quad \text{(连续)}$$

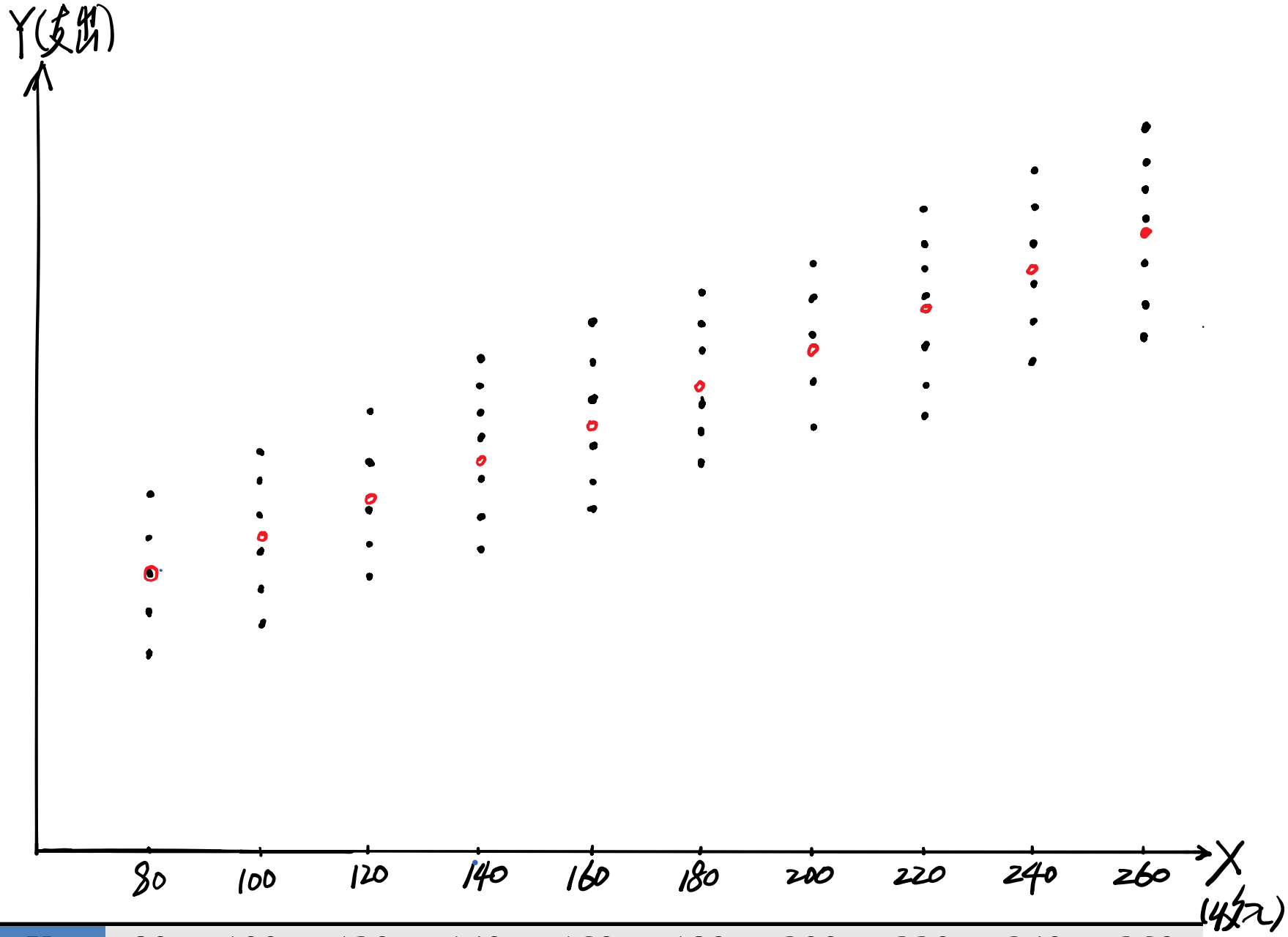
无条件期望和无条件概率

	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	$E(Y)=\sum_{i=1}^{N} Y_i \cdot p(Y_i)=\sum_{i=1}^{60}\left[55*\frac{1}{60}+60*\frac{1}{60}+\cdots+185*\frac{1}{60}+191*\frac{1}{60}\right]=\frac{1}{60}\sum_{i=1}^{N} Y_i=\frac{7272}{60}=121.2$									

条件期望和条件概率

	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	

$$\mathrm{E}(Y \mid X=80)=\sum_{i=1}^{N} Y_i \cdot p(Y_i \mid X=80)=55*\frac{1}{5}+60*\frac{1}{5}+65*\frac{1}{5}+70*\frac{1}{5}+75*\frac{1}{5}=\frac{325}{5}=65$$



X	80	100	120	140	160	180	200	220	240	260
E(Y|X)	65	77	89	101	113	125	137	149	161	173

无条件均值和条件均值

- **无条件均值：**不受X变量取值影响下，观测变量Y的均值。

$$\bar{Y} = 7272/60 = 121.2$$

- **条件均值：**在给定变量X的取值条件下，观测变量Y的均值。

$$(\bar{Y} \mid X = 80) = 65$$

条件期望、条件概率和条件均值
——计算和比较

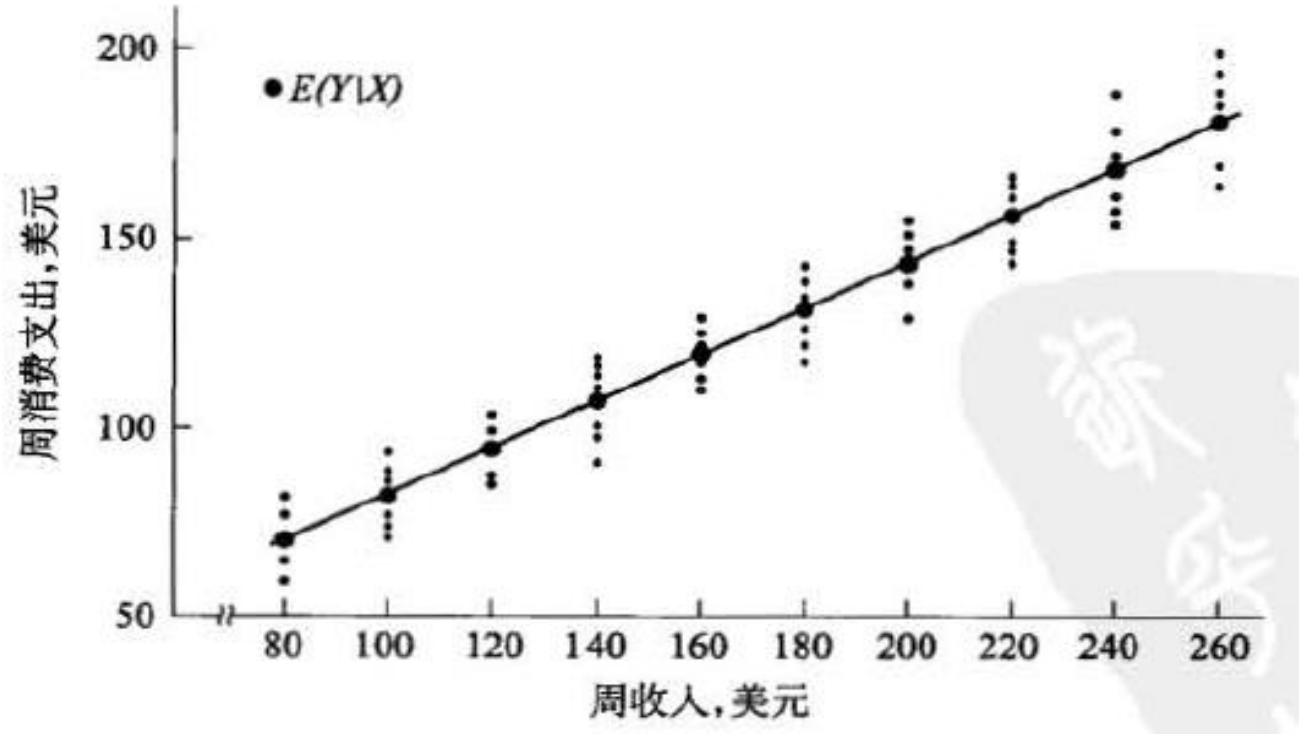
X，每周家庭收入（美元）										
Y \ X	80	100	120	140	160	180	200	220	240	260
Y，每周家庭消费支出	1/60	1/60	1/60	1/60	1/60	1/60	1/60	1/60	1/60	1/60
	1/60	1/60	1/60	1/60	1/60	1/60	1/60	1/60	1/60	1/60
	1/60	1/60	1/60	1/60	1/60	1/60	1/60	1/60	1/60	1/60
	1/60	1/60	1/60	1/60	1/60	1/60	1/60	1/60	1/60	1/60
	1/60	1/60	1/60	1/60	1/60	1/60	1/60	1/60	1/60	1/60
	—	1/60	—	1/60	1/60	1/60	—	1/60	1/60	1/60
	—	—	—	1/60	—	—	—	1/60	—	1/60
合计	325	462	445	707	678	750	685	1043	966	1211
E(Y|X)	65	77	89	101	113	125	137	149	161	173
$\bar{Y}$ | X	65	77	89	101	113	125	137	149	161	173

$$E(Y)=\sum_{i=1}^{N} Y_i \cdot p(Y_i)=\sum_{i=1}^{60}\left[55*\frac{1}{60}+60*\frac{1}{60}+\cdots+185*\frac{1}{60}+191*\frac{1}{60}\right]=\frac{1}{60}\sum_{i=1}^{N} Y_i=\frac{7272}{60}=121.2$$

总体回归线
——Population Regression Line(PRL)

- **总体回归线或总体回归曲线：**
  - 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。

Figure 2-2：



定义1：

思考：

总体回归函数
——Population Regression Function(PRF)

定义：

- **总体回归函数（PRF）：**它是总体回归线(PRL)的数学函数形式。
  - 因为总体回归线(PRL)是$E(Y|X_i)$的轨迹连成的线，所以**总体回归函数(PRF)**具有如下形式：

（式2.2.1）（PRF）

$$E(Y \mid X_i)=f(X_i)$$

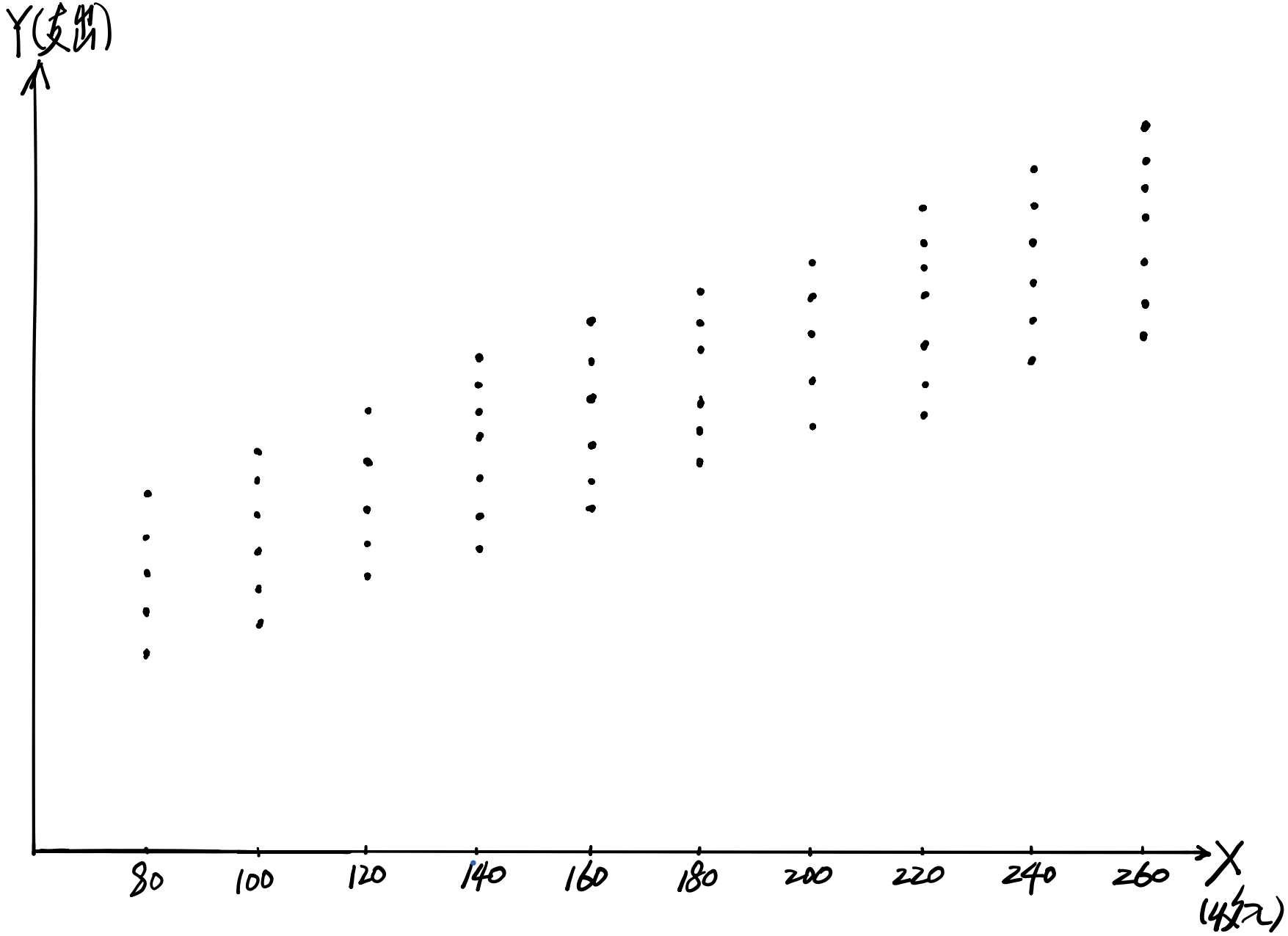
  - 对于线性总体回归函数，即有如下形式：

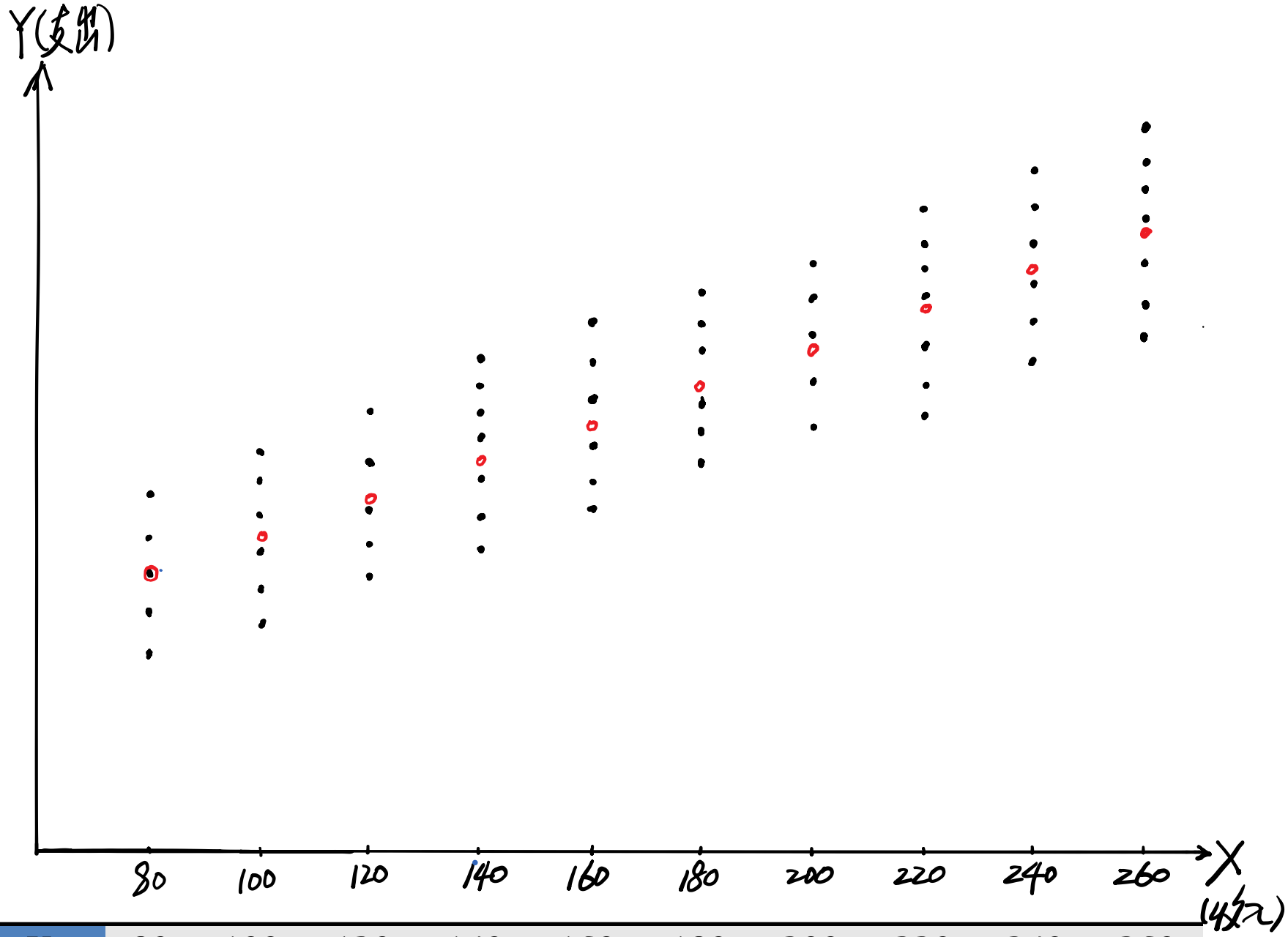
（式2.2.2）(线性PRF)

$$E(Y \mid X_i)= f(X_i)=\beta_1+\beta_2 X_i$$

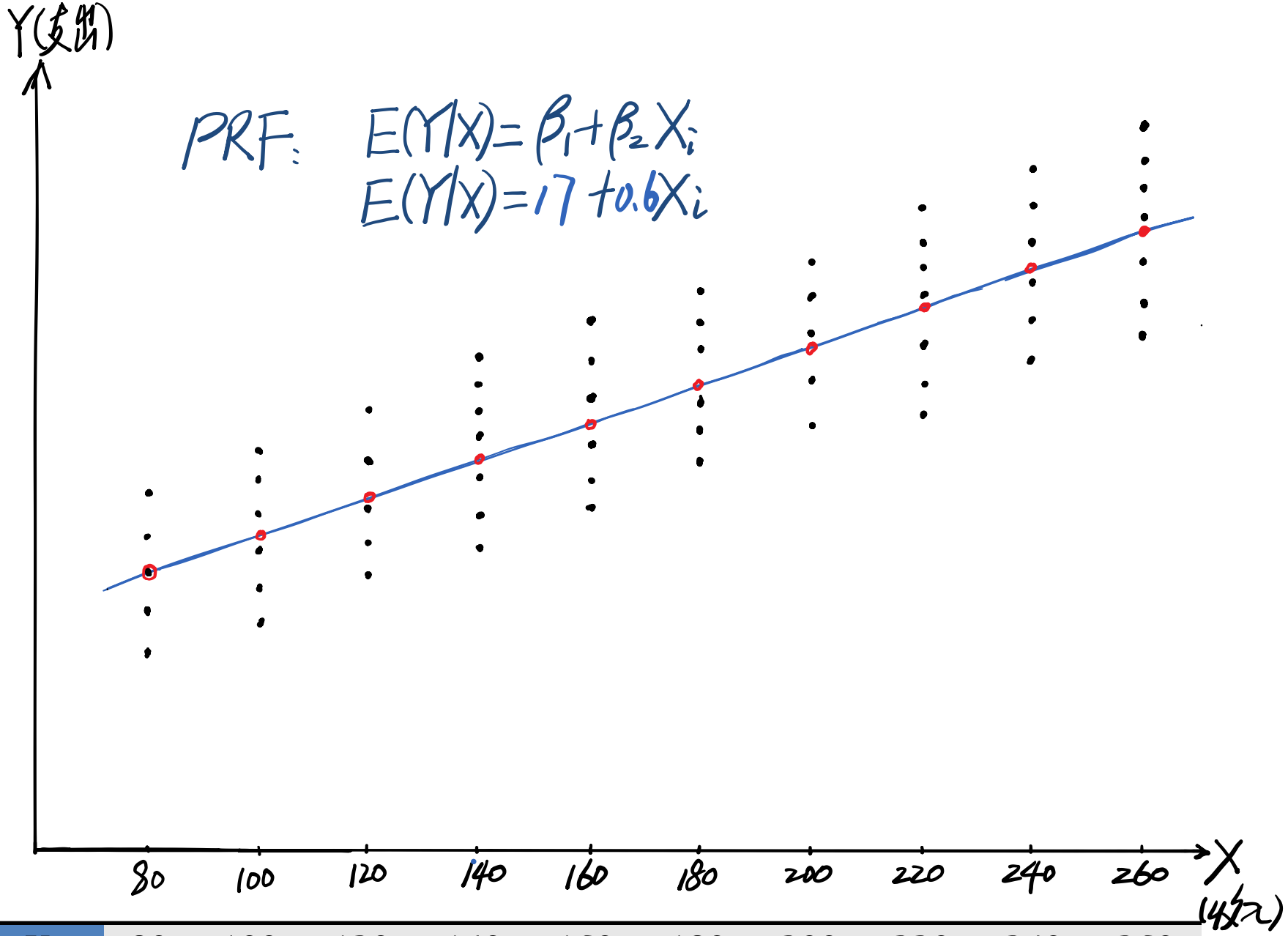
注意：

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


Y(支出)
80
100
120
140
160
180
200
220
240
260
X
(收入)



X	80	100	120	140	160	180	200	220	240	260
E(Y|X)	65	77	89	101	113	125	137	149	161	173



X	80	100	120	140	160	180	200	220	240	260
E(Y|X)	65	77	89	101	113	125	137	149	161	173

总体回归函数的随机设定
——总体回归模型（PRM）

定义1：

- **随机干扰项**或*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Yi与条件期望的离差。

$$u_i = Y_i - E(Y \mid X_i)$$

- **总体回归模型(PRM)**（或总体回归函数的随机设定）

（式2. 4. 1）
（总体回归模型）

$$Y_i = E(Y \mid X_i) + u_i$$

特定家庭的支出 = 系统性成分(systematic) / 确定性成分(deterministic) + 随机成分(stochastic) / 非系统性(nonsystematic) 成分

对总体回归模型（式2. 4. 1）两边取期望（基于X_i），得：

$$\begin{aligned} E(Y_i \mid X_i) &= E[E(Y \mid X_i)] + E(u_i \mid X_i) \\ &= E(Y \mid X_i) + E(u_i \mid X_i) \end{aligned}$$

$\because E(Y_i \mid X_i) = E(Y \mid X_i)$

$\therefore E(u_i \mid X_i) = 0$ （式2. 4. 4）

问1：这意味着总体回归函数和总体回归模型有什么关系？

随机干扰项的来源

思考：

思考*：

- **理论的抽象和简化：**理论只关注主要变量，其它变量的影响被刻意简化和忽略，导致其影响被人为归入μ_i。（如凯恩斯的绝对收入假说，只强调当期收入对消费的影响，忽略了其他因素）
- **数据的欠缺：**可能知道被忽略的变量，但不能得到这些变量的数据。（如家庭财富的数据）
- **主要变量与其它变量：**其它变量全部或其中一些合起来影响还是很小的。（如子女、教育等）
- **人类行为的内在随机性。**（客观固有的随机性）
- **变量的**测量误差（如弗里德曼的持久收入的度量）
- **建模的精简原则：遵循从简单到一般的建模思路时，**为了保持一个简单的回归模型，可能导致部分变量被忽略。
- **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（非线性关系的线性模型设定）

{理解总体回归函数和总体回归模型}

- 家庭周收入和周支出案例：
  - 总体回归函数PRF为：

$$E(Y \mid X_i) = f(X_i) = \beta_1 + \beta_2 X_i$$

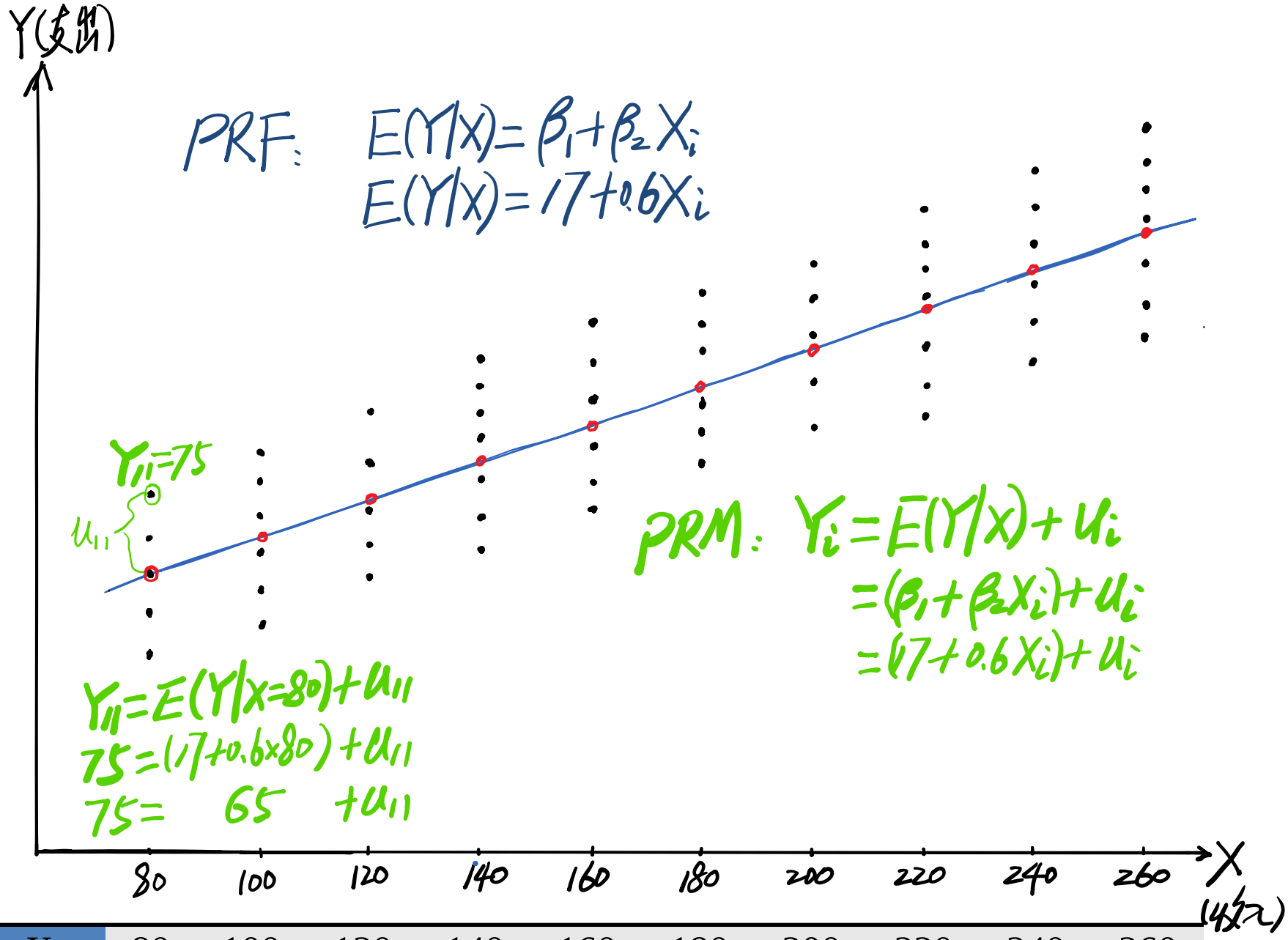
（式2.4.2）
(线性总体回归模型)

  - 总体回归模型（PRM）为：

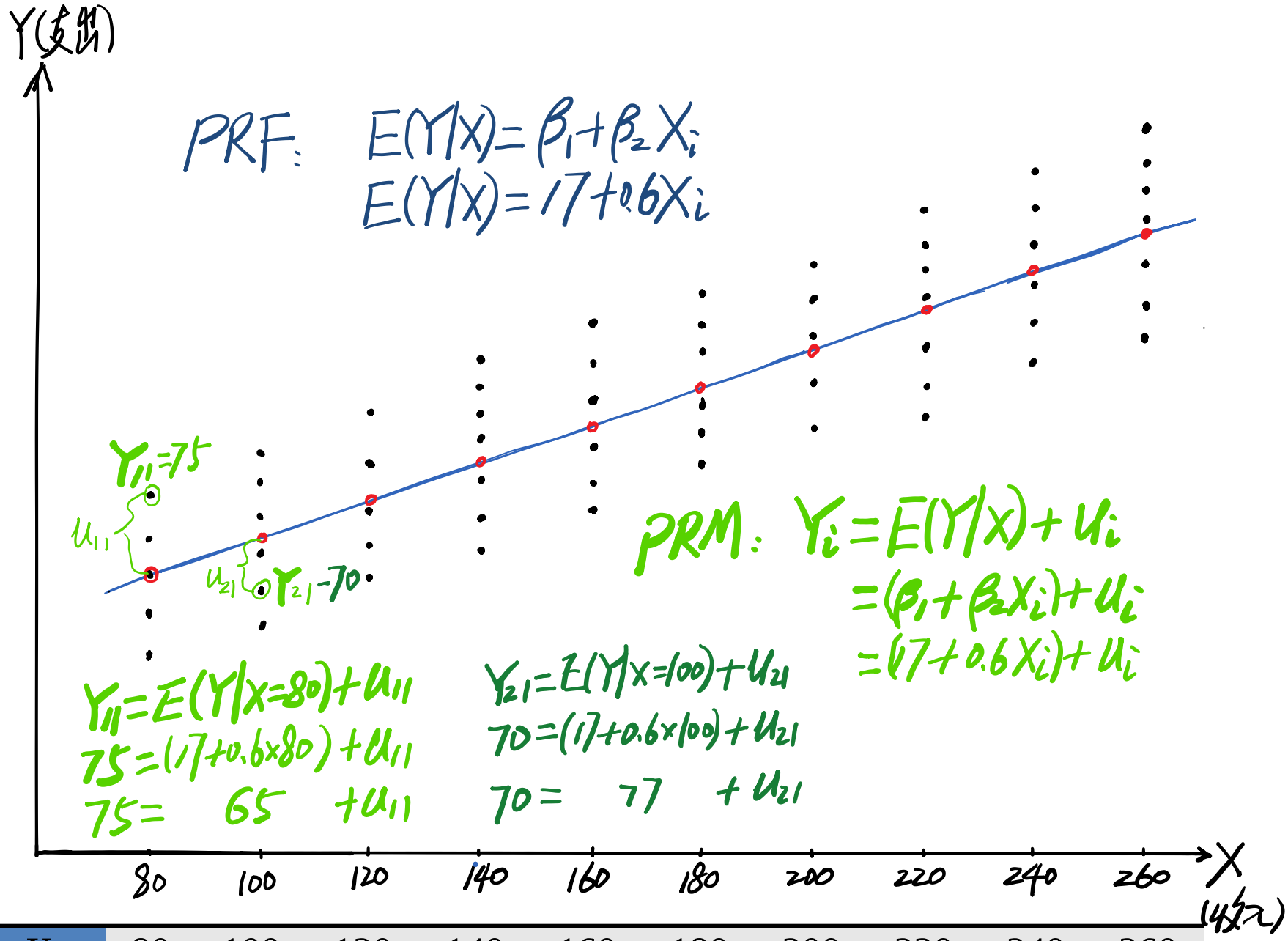
$$Y_i = E(Y \mid X_i) + u_i = \beta_1 + \beta_2 X_i + \mu_i$$

  - 若给定X_i=80美元，5个家庭的真实消费支出为：

$$\begin{aligned} Y_1 &= 55 = \beta_1 + \beta_2 80 + \mu_1 \\ Y_2 &= 60 = \beta_1 + \beta_2 80 + \mu_2 \\ Y_3 &= 65 = \beta_1 + \beta_2 80 + \mu_3 \\ Y_4 &= 70 = \beta_1 + \beta_2 80 + \mu_4 \\ Y_5 &= 75 = \beta_1 + \beta_2 80 + \mu_5 \end{aligned}$$



X	80	100	120	140	160	180	200	220	240	260
E(Y|X)	65	77	89	101	113	125	137	149	161	173



X	80	100	120	140	160	180	200	220	240	260
E(Y|X)	65	77	89	101	113	125	137	149	161	173

§ 2.2.3 “线性” 一词的含义

- **情形1：对变量为线性**
  - Y的条件期望值是Xi的线性函数，即

（式2.2.1）

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

- **情形2：对参数为线性**
  - “线性”回归一词总是指对参数 β 为线性的一种回归；对解释变量X则可以是或不是线性的。（与以往的线性概念有所不同）

（式2.2.2）

$$E(\mathrm{Y} \mid \mathrm{X}i) = \beta_1 + \beta_2 X_i^2; \quad \mathrm{E}(\mathrm{Y} \mid \mathrm{X}i) = \beta_1 + \beta_2 (\frac{1}{X_i});$$

提问：

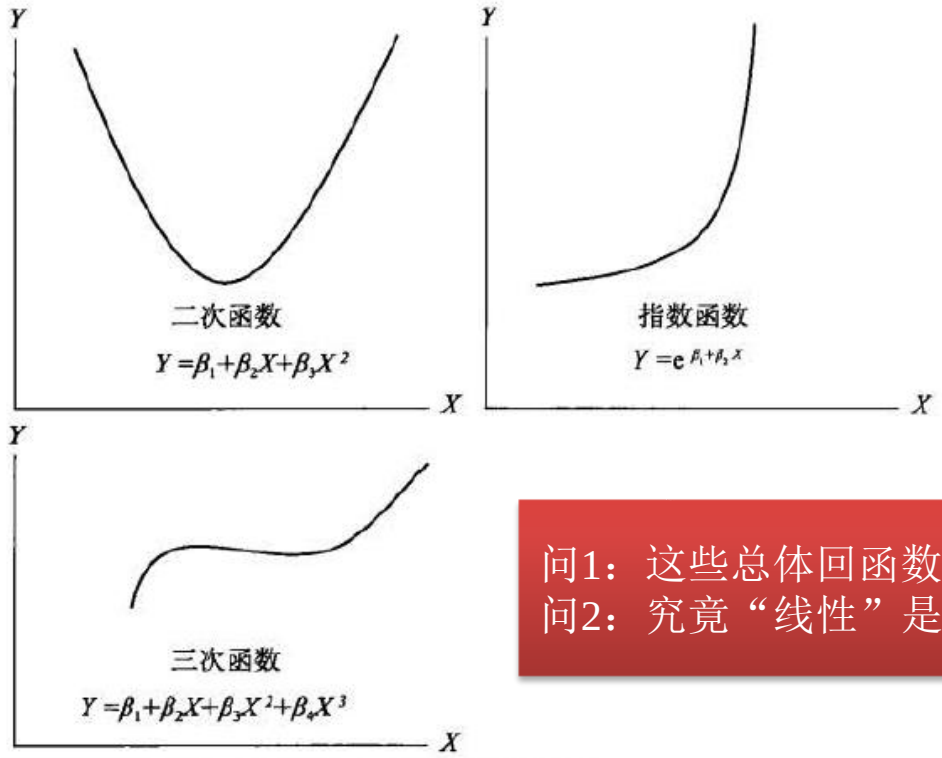
$$\mathrm{E}(\mathrm{Y} \mid \mathrm{X}_i) = \beta_1 + \beta_2^2 X_i; \quad \mathrm{E}(\mathrm{Y} \mid \mathrm{X}_i) = \beta_1 + \beta_2 X_i;$$

$$\mathrm{E}(\mathrm{Y} \mid \mathrm{X}_i) = \beta_1 + \beta_2 \ln(X_i); \quad \mathrm{E}(\mathrm{Y} \mid \mathrm{X}_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2;$$

$$\ln(\mathrm{E}(\mathrm{Y} \mid \mathrm{X}_i)) = \beta_1 - \beta_2 (\frac{1}{X_i});$$

深入理解“线性”回归的内涵*

思考（提问）



问1：这些总体回函数为什么也能称为“线性”？
问2：究竟“线性”是指什么？（common sense？）

模型对参数为线性？	模型对变量为线性？	
	是	不是
是	LRM	LRM
不是	NLRM	NLRM

注：LRM=线性回归模型；
NLRM=非线性回归模型。

深入理解“线性”回归的内涵*

思考（提问）

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \varepsilon_i \rightarrow quadratic$$

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \beta_4 x_i^3 + \varepsilon_i \rightarrow cubic$$

$$y_i = \beta_1 + \beta_2 \left(\frac{1}{x_i} \right) + \varepsilon_i \rightarrow reciprocal$$

$$y_i = \beta_1 + \beta_2 \ln x_i + \varepsilon_i \rightarrow semilogarithmic$$

$$\ln y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \rightarrow inverse\ semilogarithmic$$

$$\ln y_i = \beta_1 - \beta_2 \left(\frac{1}{x_i} \right) + \varepsilon_i \rightarrow logarithmic\ reciprocal$$

$$\ln y_i = \ln \beta_1 + \beta_2 \ln x_i + \varepsilon_i \rightarrow logarithmic\ or\ double\ logarithmic$$

$$y_i = e^{\beta_1 + \beta_2 x_i + \varepsilon_i} \rightarrow exponential$$

$$y_i = \frac{1}{1 + e^{\beta_1 + \beta_2 x_i + \varepsilon_i}} \rightarrow logistic\ (probability)\ distribution\ function$$

$$y_i = \beta_1 + (0.75 - \beta_1) e^{-\beta_2 (x_i - 2)} + \varepsilon_i$$

$$y_i = \beta_1 + \beta_2^3 x_i + \varepsilon_i$$

样本回归线(SRL)和样本回归函数(SRF)

约定：

Table 2-4/5

（式2.6.1）
(线性SRF)

注意：

- **样本特征和总体特征：**
  - 总体特征值是一常数称为参数
  - 样本特征值是一变数称为估计量或统计量
- **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
  - 例如采用OLS方法对样本数据进行拟合
- **样本回归函数(Sample Regression Function，SRF):**(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：

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

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

- $\hat{Y}_i$是$E(Y|X_i)$的估计量
- $\widehat{\beta_1}$是β_1的估计量；
- $\widehat{\beta_2}$是β_2的估计量；

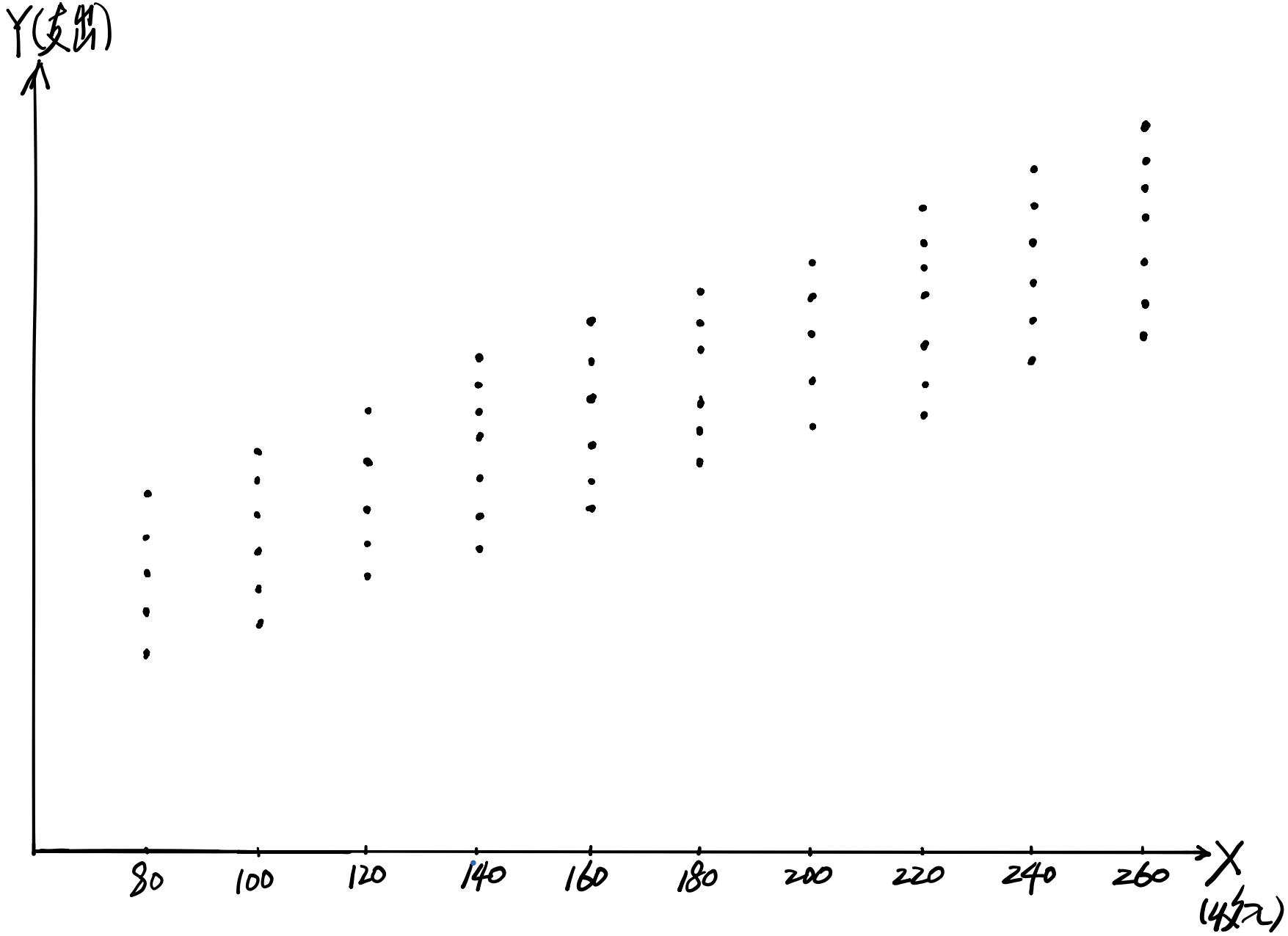
线性SRF　　线性PRF

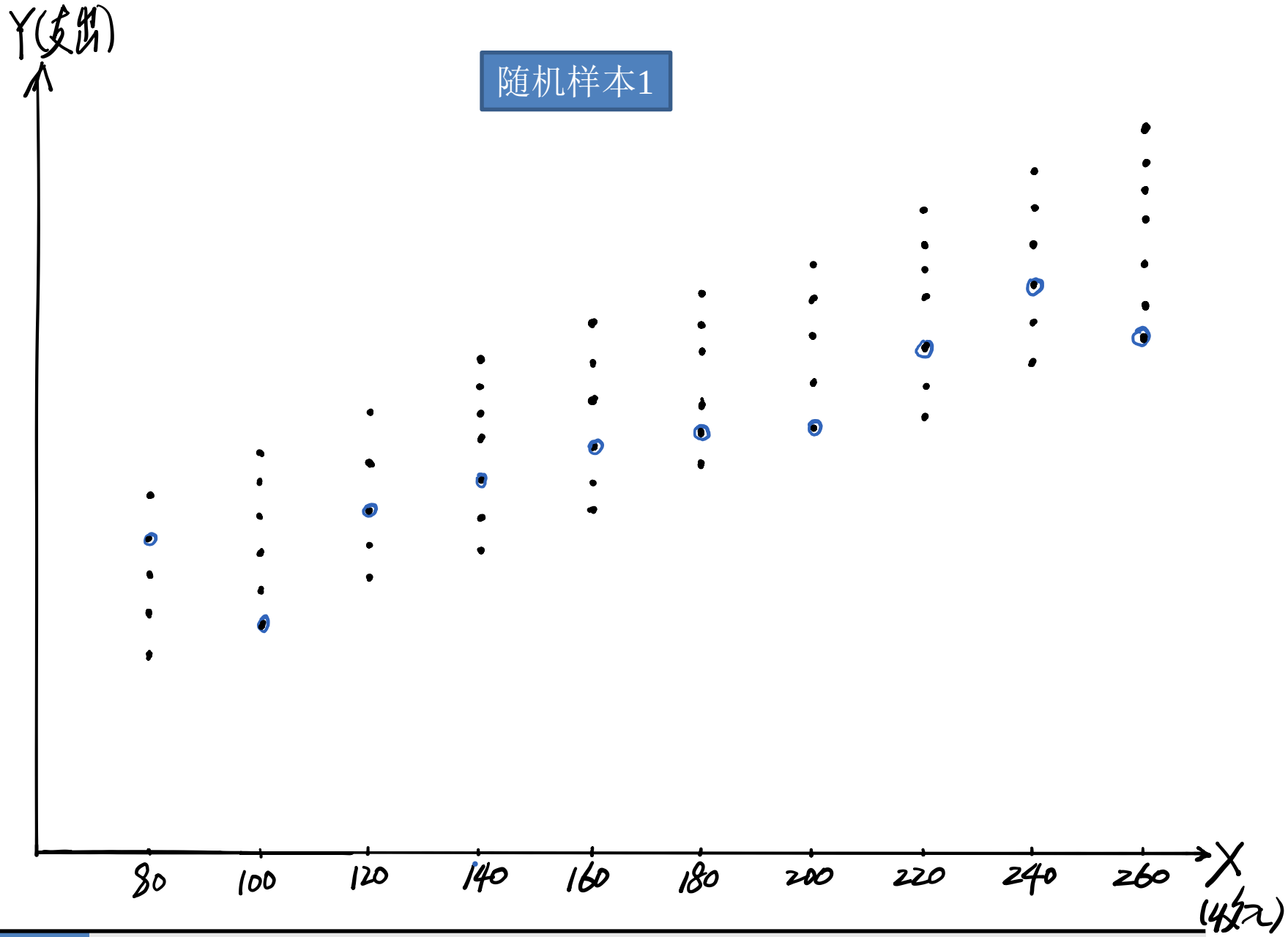
样本数据特征：
——两份随机样本

总体

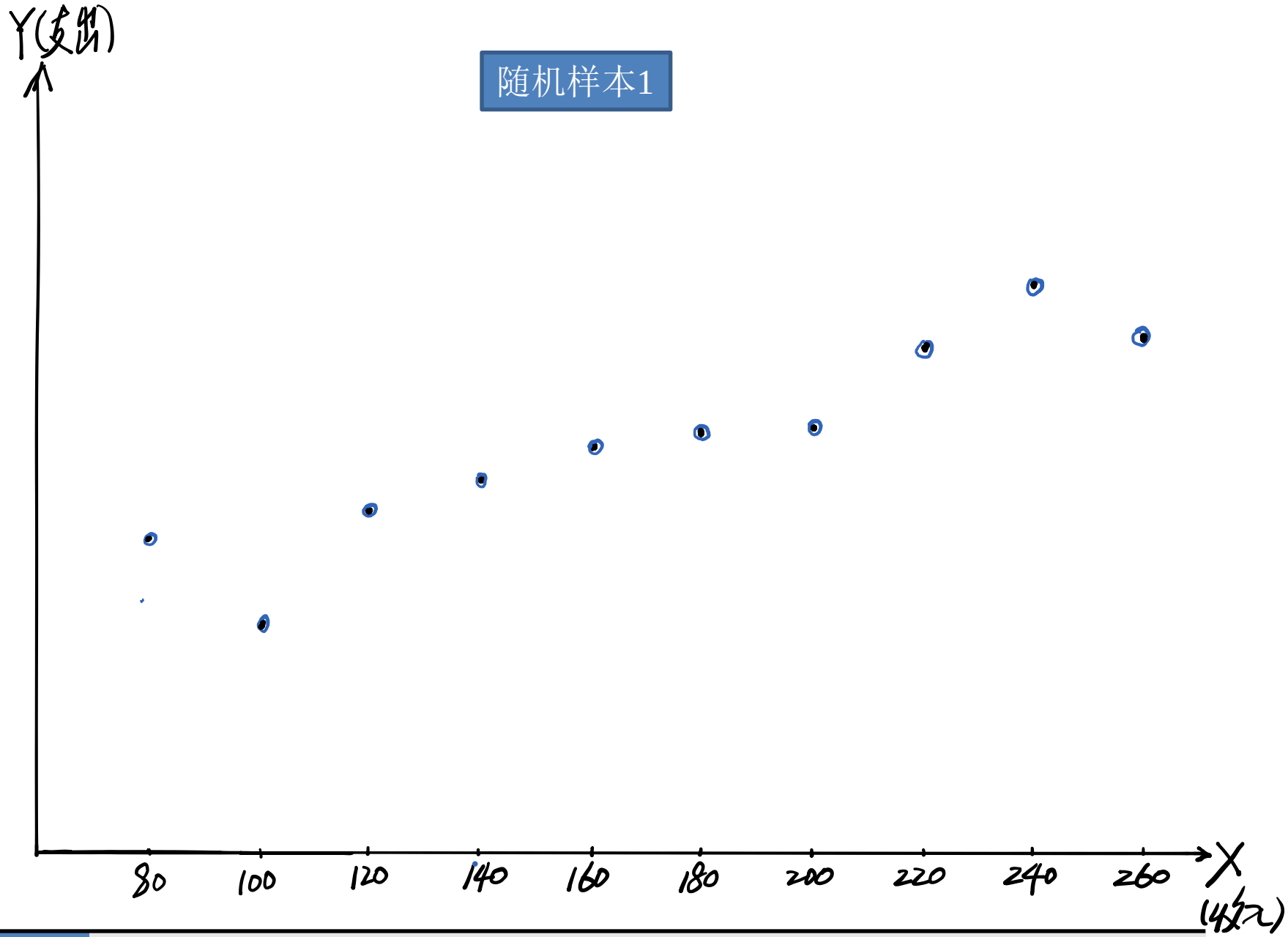
X，每周家庭收入（美元）										
X / Y	80	100	120	140	160	180	200	220	240	260
Y，每周家庭消费支出	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
	—	88	—	113	125	140	—	160	189	185
	—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

样本1	X	80	100	120	140	160	180	200	220	240	260
	Y	**70**	**65**	**90**	**95**	**110**	**115**	**120**	**140**	**155**	**150**
样本2	X	80	100	120	140	160	180	200	220	240	260
	Y	**55**	**88**	**90**	**80**	**118**	**120**	**145**	**135**	**145**	**175**

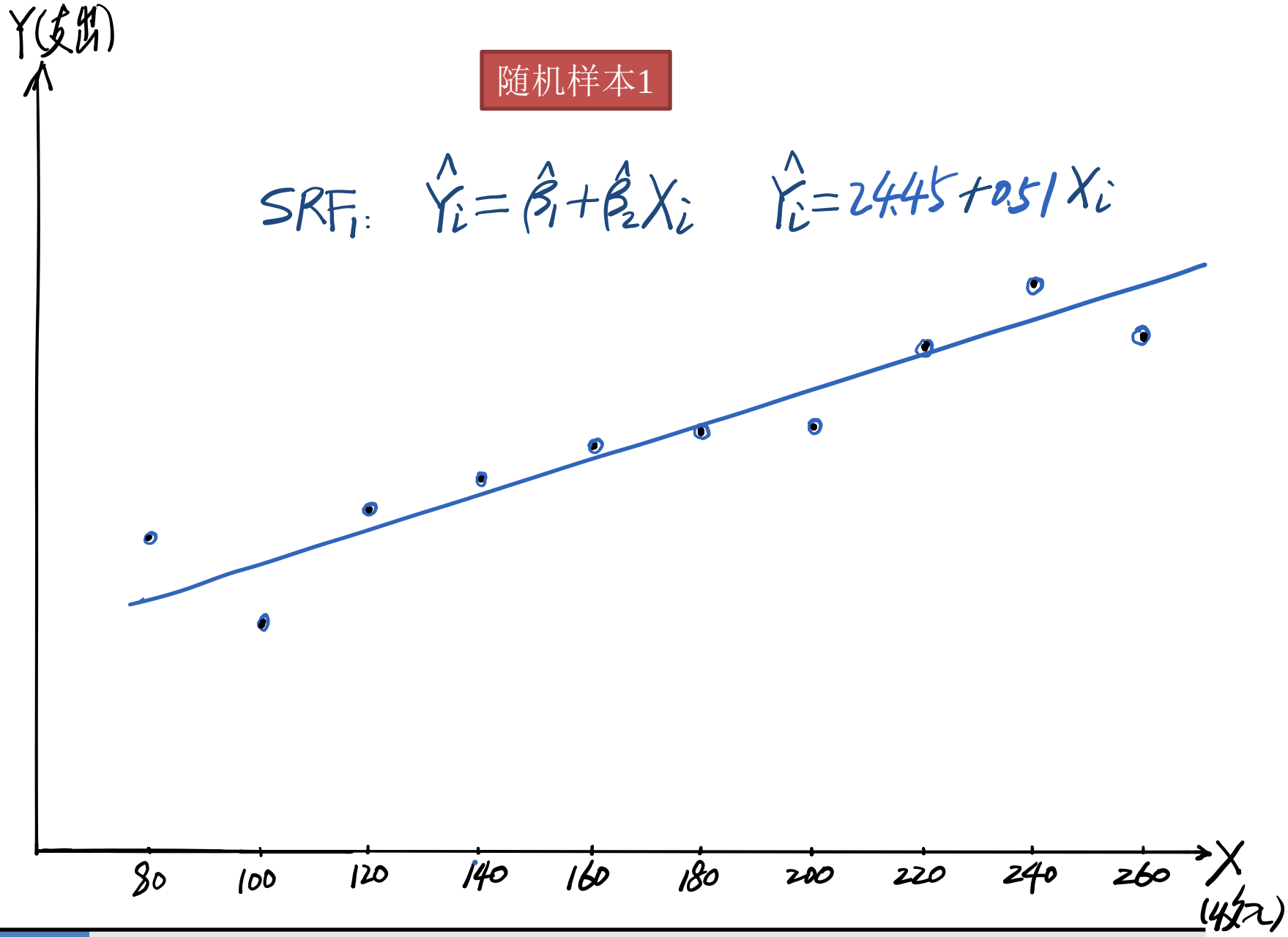

Y(支出)
80
100
120
140
160
180
200
220
240
260
X
(收入)



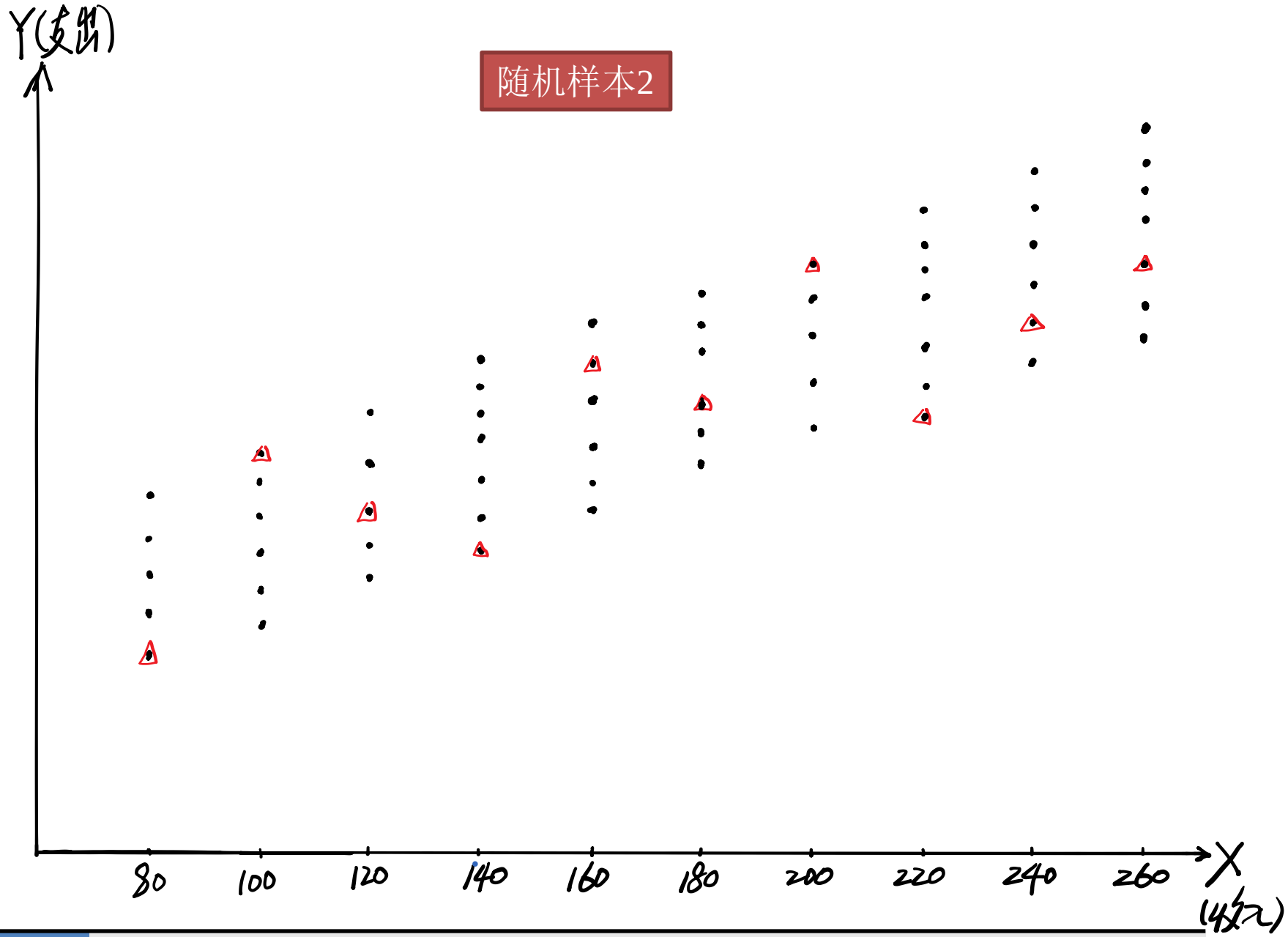
X	80	100	120	140	160	180	200	220	240	260
Y	**70**	**65**	**90**	**95**	**110**	**115**	**120**	**140**	**155**	**150**



X	80	100	120	140	160	180	200	220	240	260
Y	**70**	**65**	**90**	**95**	**110**	**115**	**120**	**140**	**155**	**150**

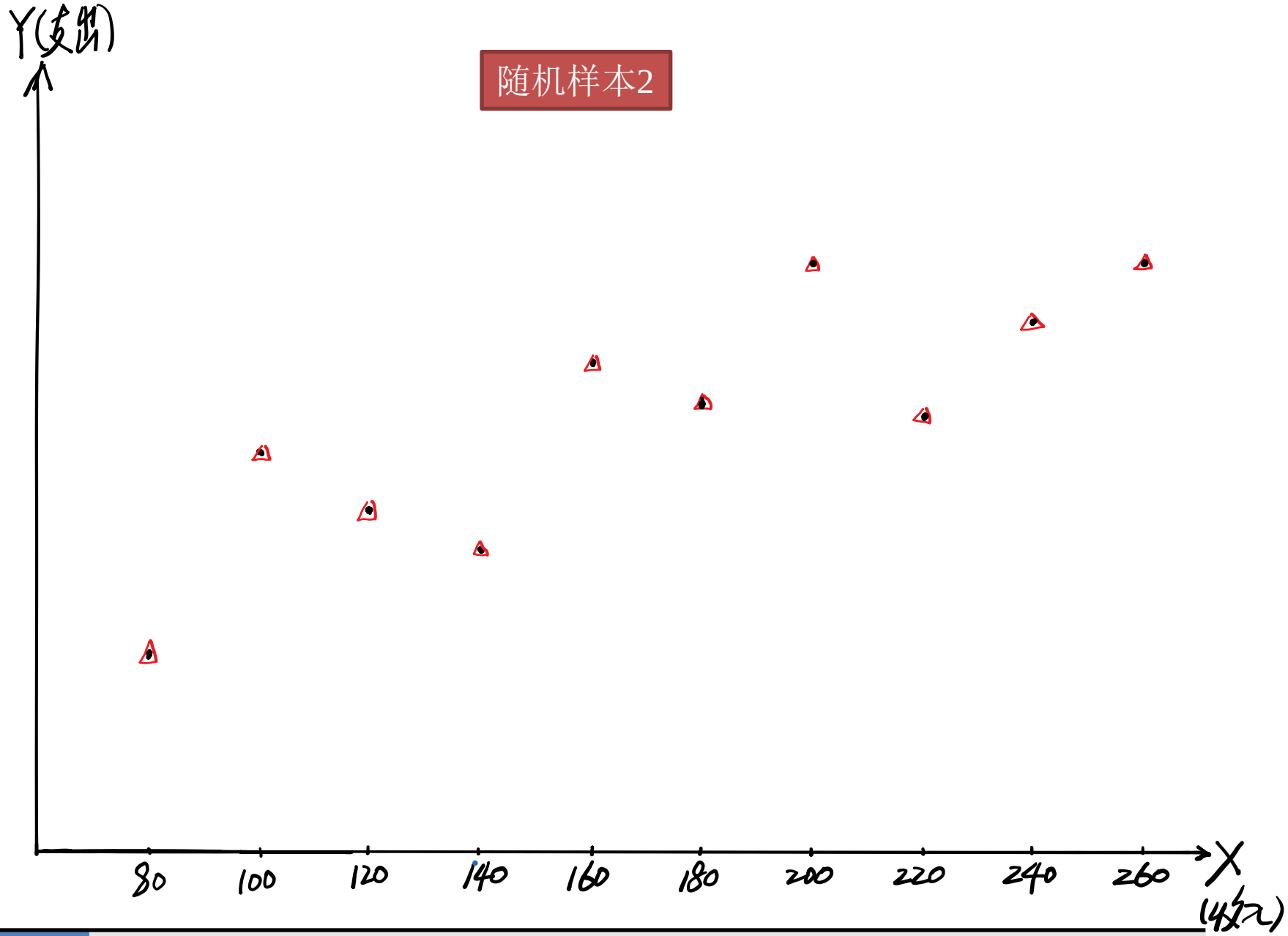


X	80	100	120	140	160	180	200	220	240	260
Y	55	88	90	80	118	120	145	135	145	175



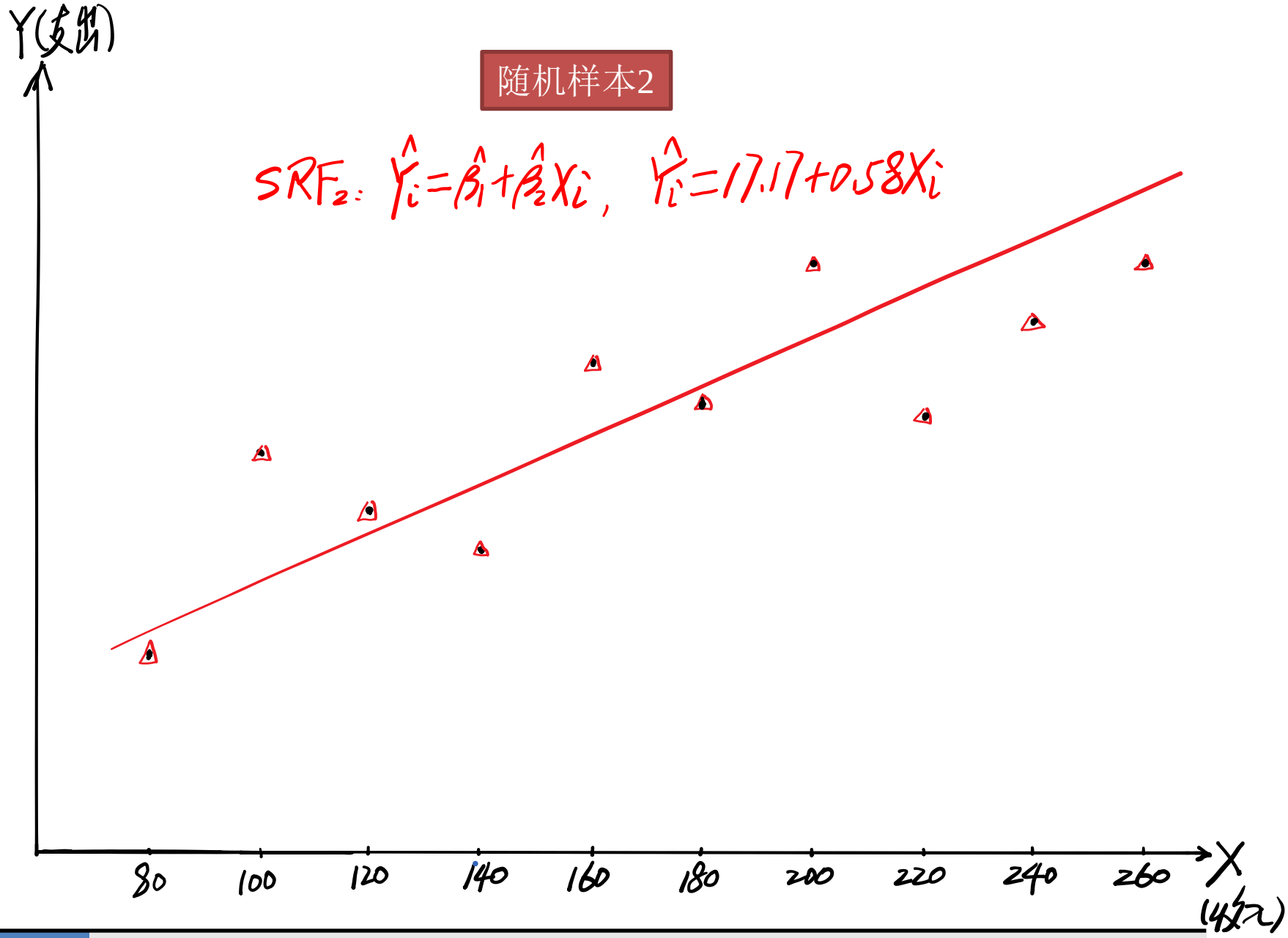
X	80	100	120	140	160	180	200	220	240	260
Y	55	88	90	80	118	120	145	135	145	175

随机样本2



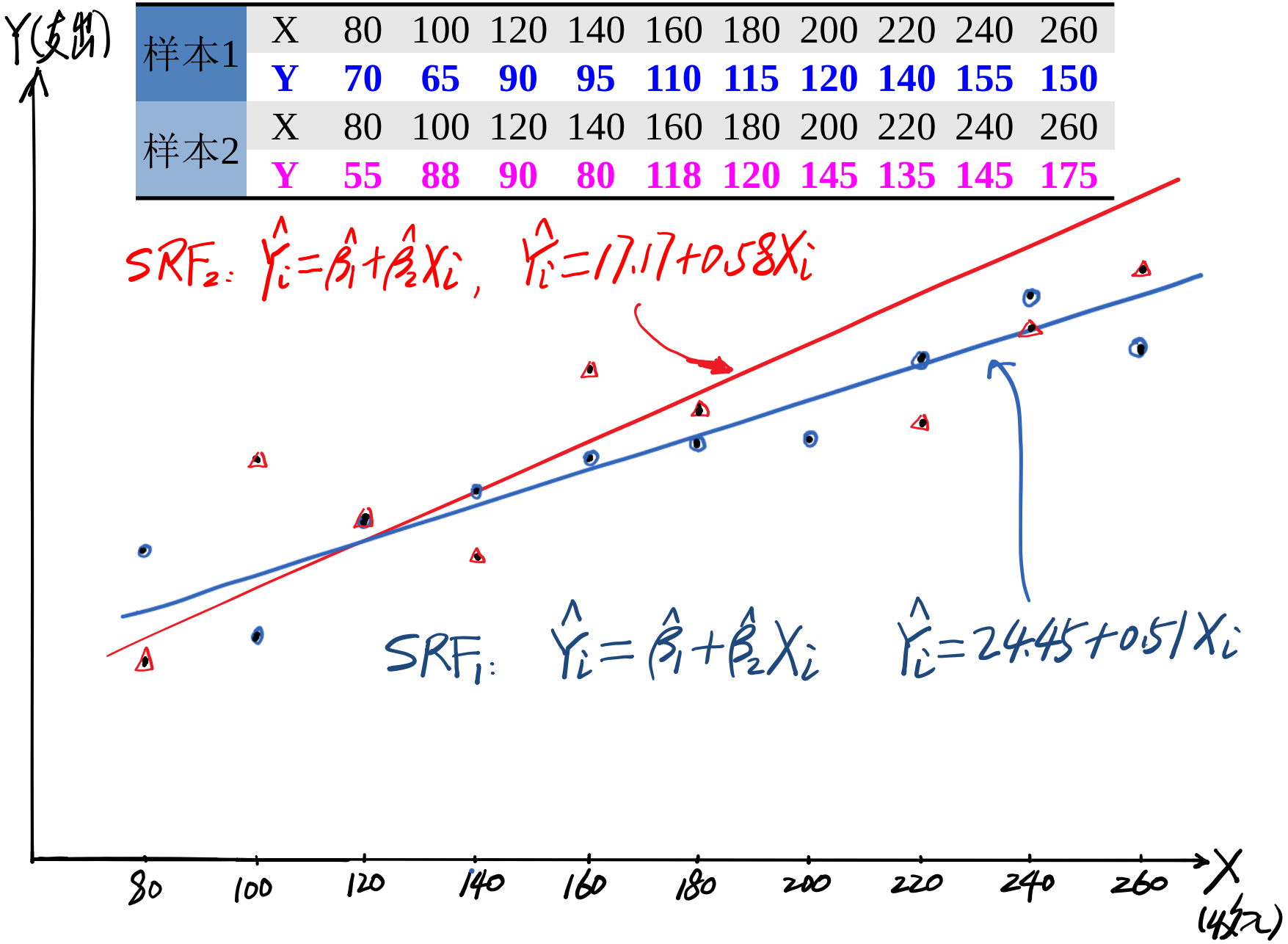
X	80	100	120	140	160	180	200	220	240	260
Y	55	88	90	80	118	120	145	135	145	175

随机样本2

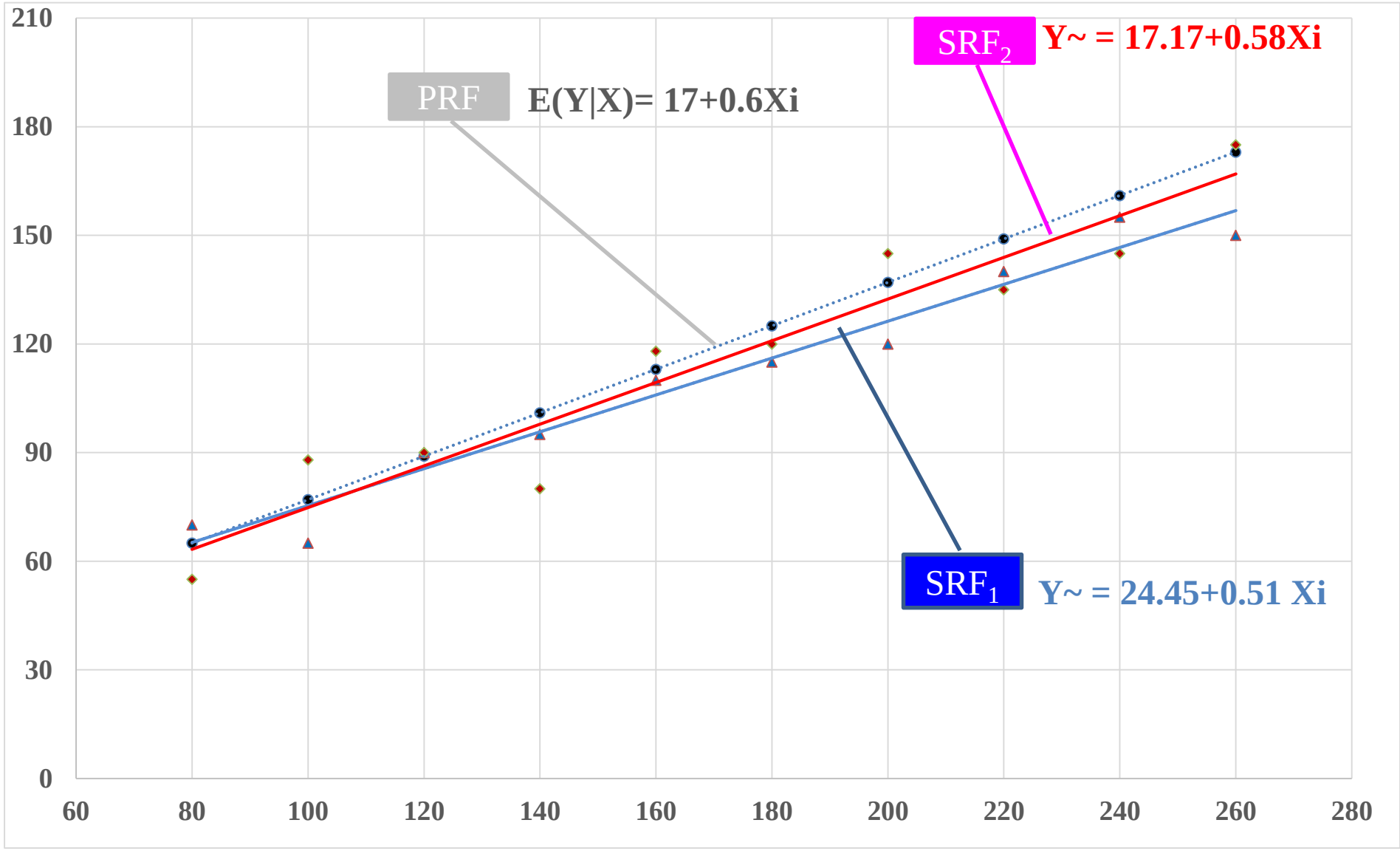


X	80	100	120	140	160	180	200	220	240	260
Y	55	88	90	80	118	120	145	135	145	175

样本1	X	80	100	120	140	160	180	200	220	240	260
	Y	**70**	**65**	**90**	**95**	**110**	**115**	**120**	**140**	**155**	**150**
样本2	X	80	100	120	140	160	180	200	220	240	260
	Y	**55**	**88**	**90**	**80**	**118**	**120**	**145**	**135**	**145**	**175**



样本特征能不能/如何代表总体特征？



样本回归函数的随机形式
——样本回归模型（SRM）

- **残差（Residual）：** 是样本回归函数与Y的样本观测值之间的离差。

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

- **样本回归模型**（或样本回归函数的随机形式）：

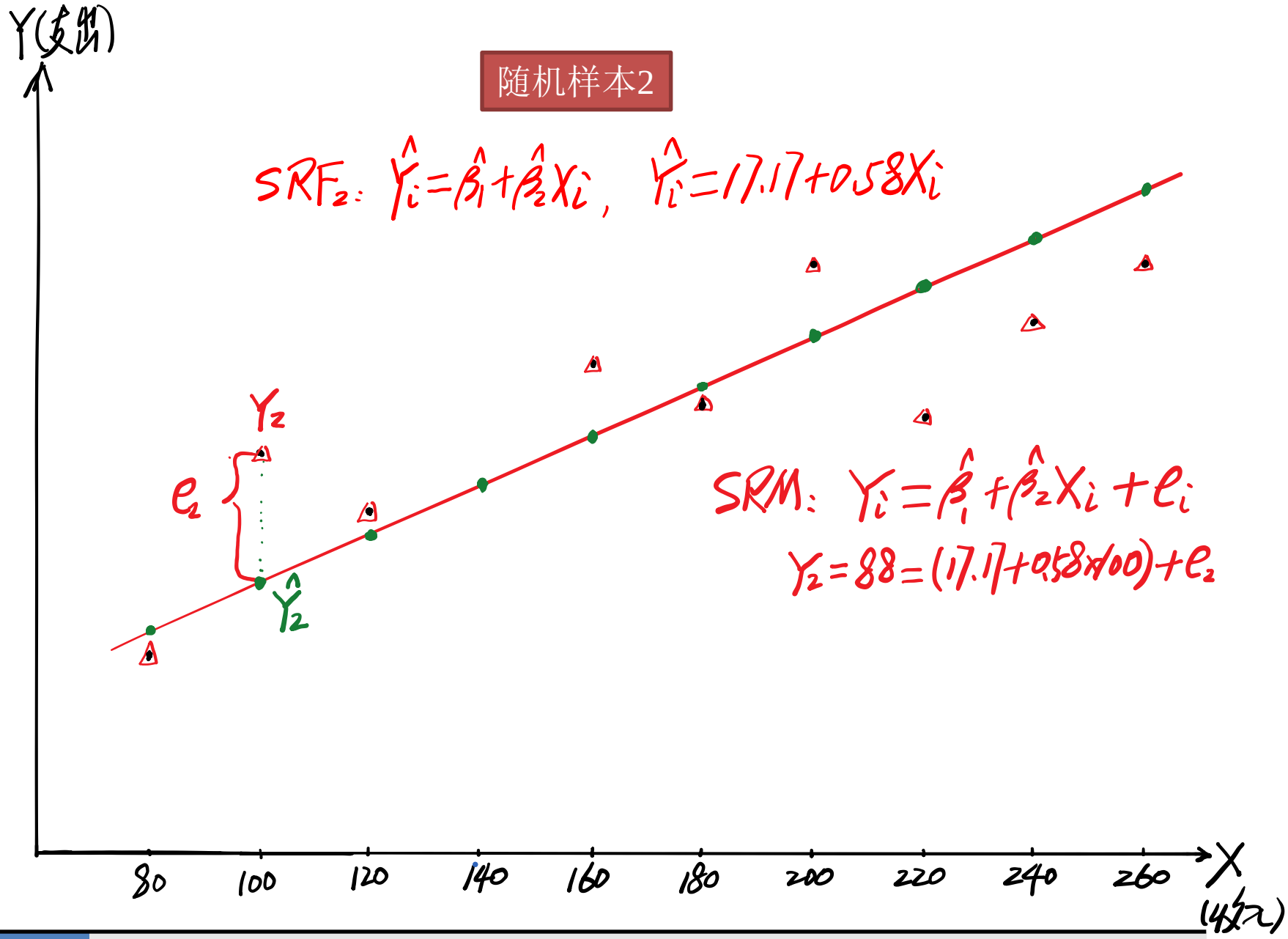
$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$$

线性SRM

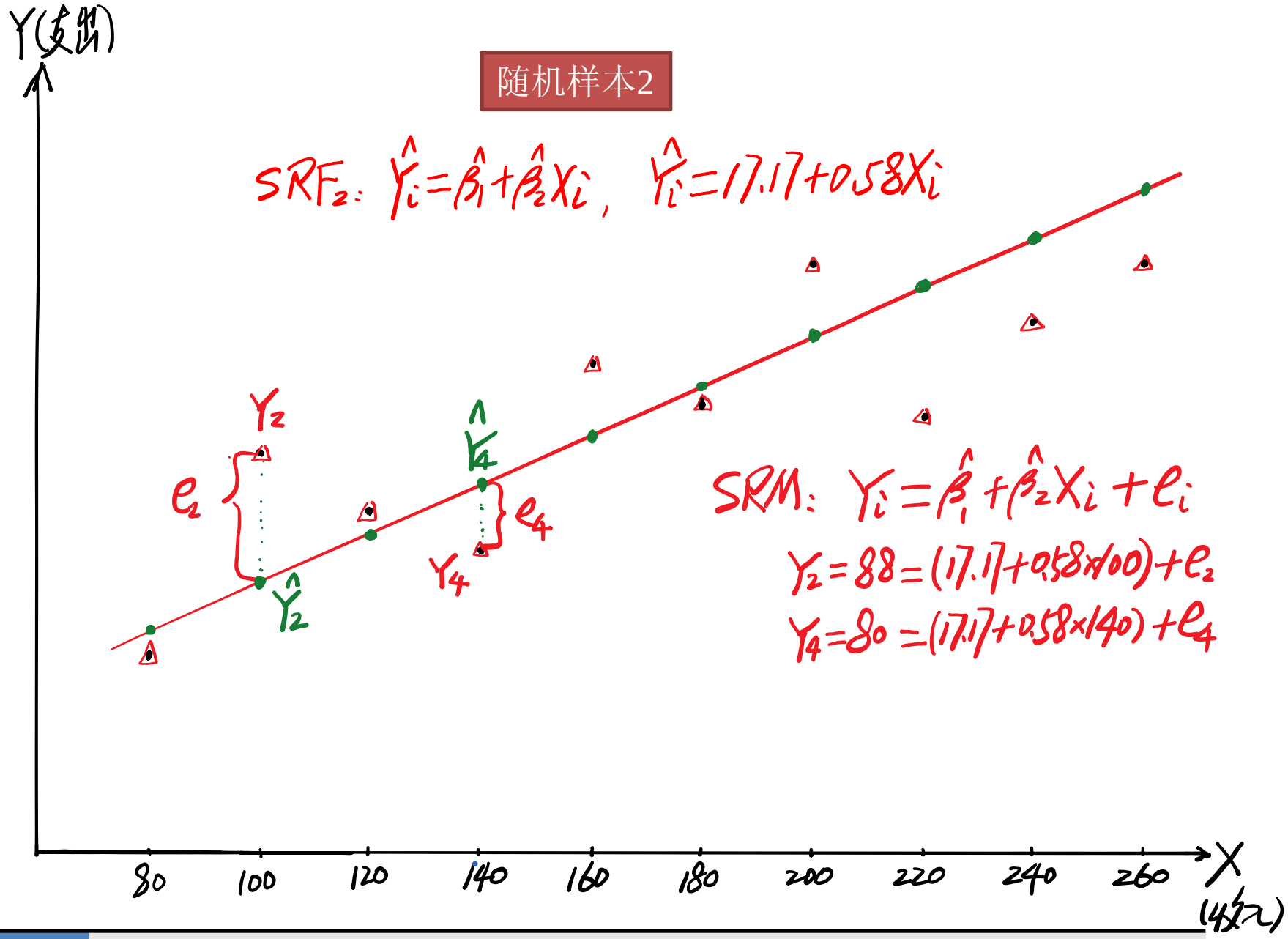
$$Y_i = \beta_1 + \beta_2 X_i + \mu_i$$

线性PRM

e_i是对μ_i的一个估计量



X	80	100	120	140	160	180	200	220	240	260
Y	55	88	90	80	118	120	145	135	145	175



X	80	100	120	140	160	180	200	220	240	260
Y	55	88	90	80	118	120	145	135	145	175

样本回归与总体回归的比较

Figure 2-5

注意：

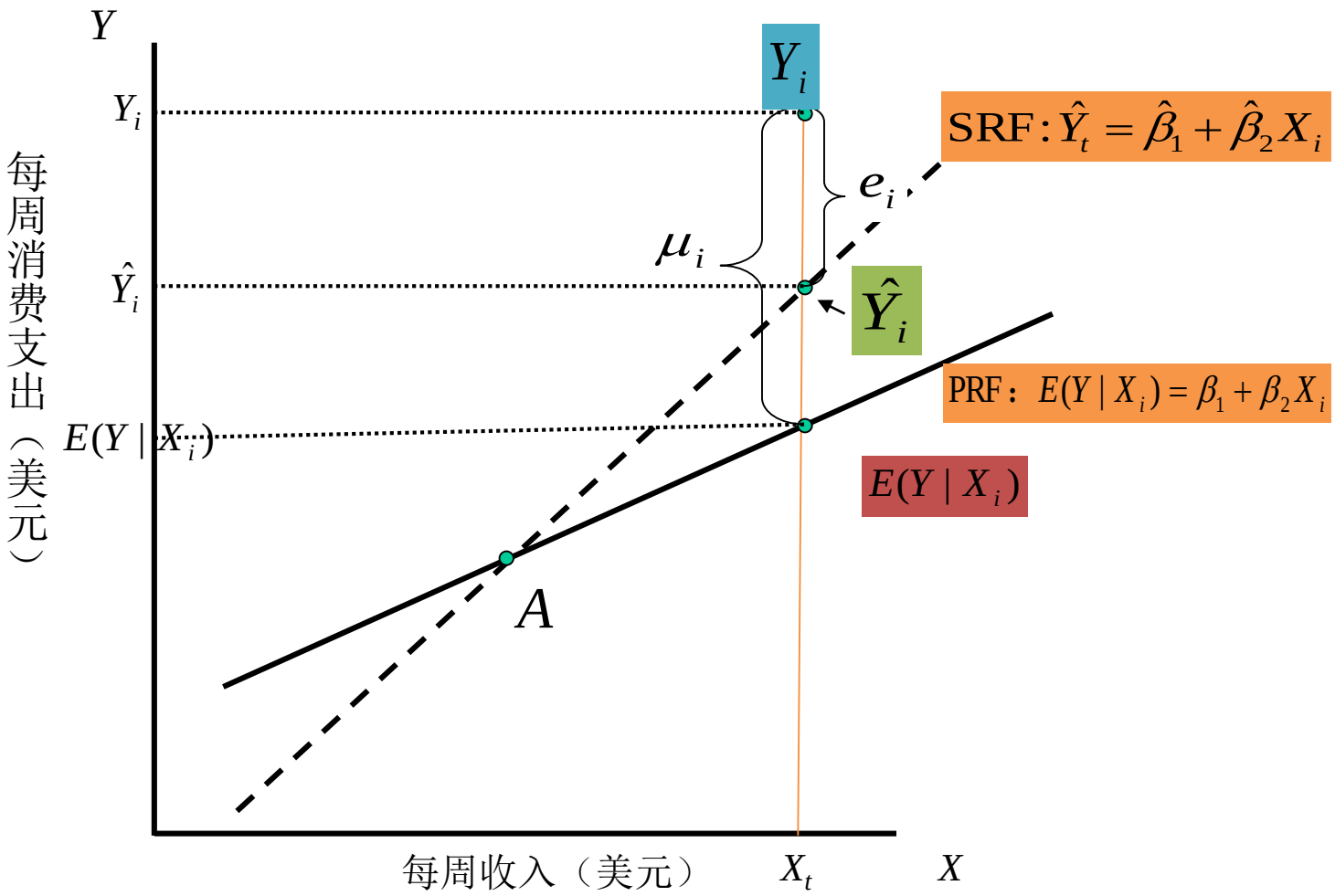


图2-5 样本与总体回归线的比较

样本回归与总体回归的比较

(PRF) $E(Y \mid X_i) = f(X_i)$ 总体回归函数

(L-PRF) $E(Y \mid X_i) = \beta_1 + \beta_2 X_i$ 总体回归函数

(PRM) $Y_i = E(Y \mid X_i) + u_i$ 总体回归模型

(L-PRM) $Y_i = \beta_1 + \beta_2 X_i + \mu_i$ 线性总体回归模型

(SRF) $\hat{Y}_i = f(X_i)$ 样本回归函数

(L-SRF) $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$ 线性样本回归函数

(SRM) $Y_i = f(X_i) + e_i$ 样本回归模型

(L-SRM) 线性回归模型

思考：

问1：SRF是不是对PRF 的一个近似？

问2： β_1和 β_2能不能算出来？ $\widehat{\beta_1}$ 和 $\widehat{\beta_2}$又能不能算出来？

问3：能不能设计一种规则或方法，使得这种近似是一种尽可能“接近”地近似?

§ 2.2.7 说明性例子
——Example 1： 受教育程度与时均工资

Table 2-6

表2-6 不同受教育程度的平均小时工资

分组	读书年数	时均工资$	人数
1	6	4.4567	3
2	7	5.77	5
3	8	5.9787	15
4	9	7.3317	12
5	10	7.3182	17
6	11	6.5844	27
7	12	7.8182	218
8	13	7.8351	37
9	14	11.0223	56
10	15	10.6738	13
11	16	10.8361	70
12	17	13.615	24
13	18	13.531	31
			528

此数据源自：1985年5月的美国人口普查

Figure 2-6

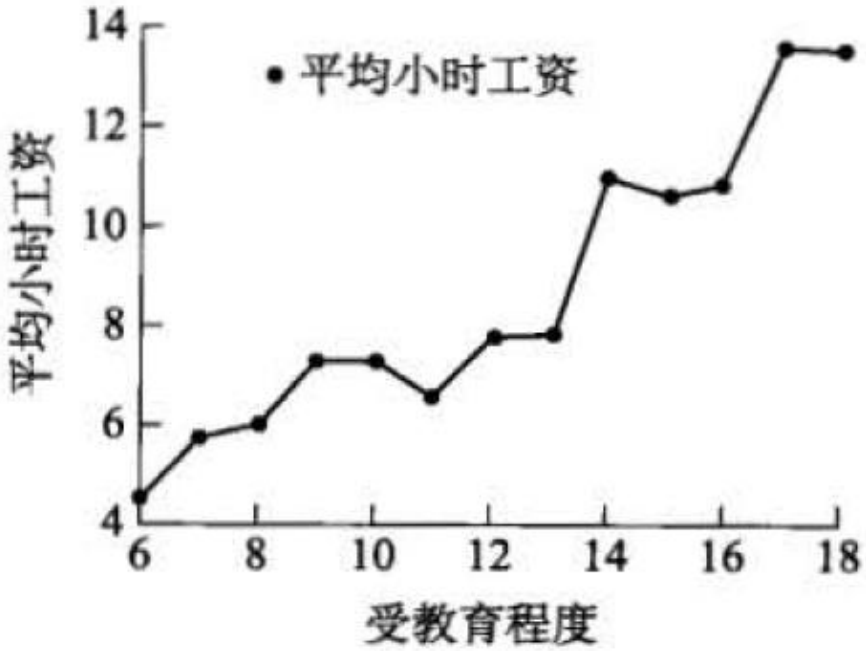


图2-6 平均小时工资与受教育程度之间的关系

思考：

Example 2： 家庭收入与考试成绩

Table 2-10

家庭收入		应试	阅读		数学		写作	
分组	组中值	人数	均值	标准差	均值	标准差	均值	标准差
< 10,000	5000	40610	427	107	451	122	423	104
10000-20000	15000	72745	453	106	472	113	446	102
20000-30000	25000	61244	454	102	465	107	444	97
30000-40000	35000	83685	476	103	485	106	466	98
40000-50000	45000	75836	489	103	486	105	477	99
50000-60000	55000	80060	497	102	504	104	486	98
60000-70000	65000	75763	504	102	511	103	493	98
70000-80000	75000	81627	508	101	516	103	498	98
80000-100000	90000	130752	520	102	529	104	510	100
>100000	150000	245025	544	105	556	107	537	103

此数据源自：据2007 年参加SAT 考试的947 347 名考生信息

表2-10 家庭收入与成绩

Figure 2-7

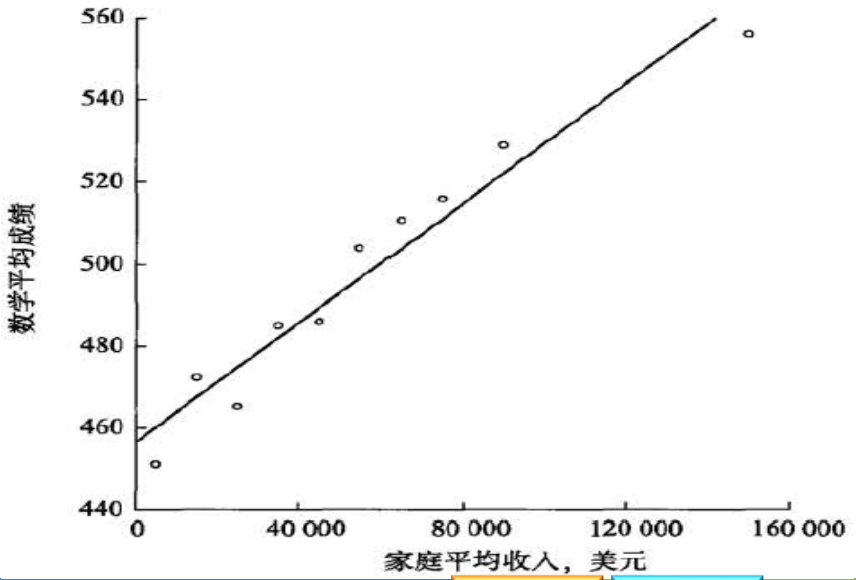


图 2-7 家庭收入与数学成绩

思考：

本节内容：

§2.3 双变量回归分析：估计问题

§ 2.3.1普通最小二乘法（OLS）
——如何构造SRF

- 主要是估计回归函数中的系数
- 有多种方法
  - 图解法
  - 最小二乘法(order lease squares, OLS)
  - 最大似然法(maximum likelihood, ML)
- 最常用的方法：最小二乘法

思考：

认识普通最小二乘法的原理

(L-PRF) $E(Y \mid X_i) = \beta_1 + \beta_2 X_i$ 线性总体回归函数

(L-PRM) $Y_i = E(Y \mid X_i) + u_i = \beta_1 + \beta_2 X_i + \mu_i$ 线性总体回归模型

(L-SRF) $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$ 线性样本回归函数

(L-SRM) $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$ 线性样本回归模型

- 因为PRF无法直接观测，只能用SRF近似替代

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

- 那么，估计值与观测值之间存在偏差(残差)：

（式3.1.1）(residual)

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

思考：

- 但是，SRF 又是怎样决定的呢?

认识普通最小二乘法的原理(1)：一个图示

Figure 3-1

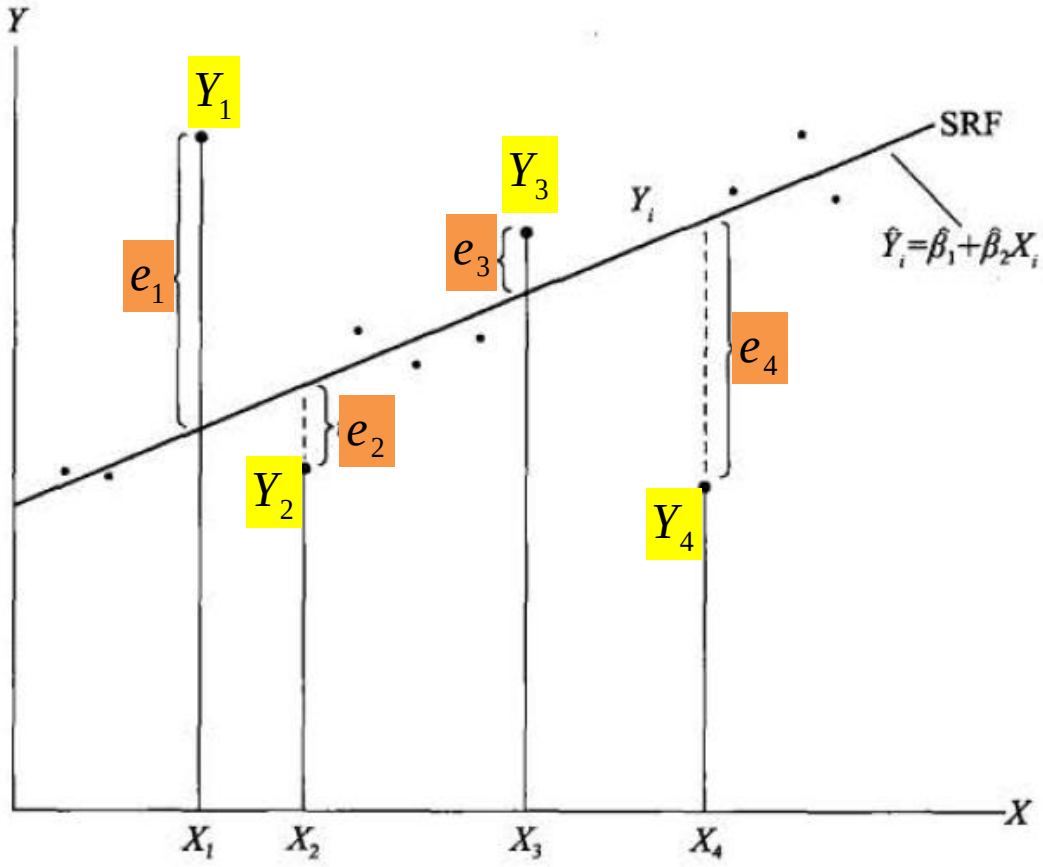


图3-1 最小二乘法的原理

原理：

$$\min_{\hat{\beta}_1,\hat{\beta}_2} S=\sum_{i=1}^{n}\hat{u}_i^2=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2=\sum_{i=1}^{n}(Y_i-(\hat{\beta}_1+\hat{\beta}_2X_i))^2=S(\hat{\beta}_1,\hat{\beta}_2)$$

认识普通最小二乘法的原理(2)：一个数值试验

- 假设存在（表3-1）所示的4组观测值$(\boldsymbol{Y_i},\boldsymbol{X_i})$；
- 假设随便猜想了如下两个SRF：

$$SRF1:\ \hat{Y}_{1i}=\hat{\beta}_1+\hat{\beta}_2X_i=1.572+1.357X_i\ ,\quad 而且\quad e_{1i}=Y_i-\hat{Y}_{1i}$$

$$SRF2:\ \hat{Y}_{2i}=\hat{\beta}_1+\hat{\beta}_2X_i=3.0+1.0X_i\ ,\quad 而且\quad e_{2i}=Y_i-\hat{Y}_{2i}$$

练习：

- 完成下表计算，并分析哪个SRF给出$(\hat{\beta}_1,\hat{\beta}_2)$要更好？

Table 3-1

Y_i	X_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)
4	1						
5	4						
7	5						
12	6						
Sum：							

思考：

用OLS方法估计总体参数β_i
——得到正规方程组

$$S(\hat{\beta}_1,\hat{\beta}_2)=\sum_{i=1}^{n}(Y_i-\hat{\beta}_1-\hat{\beta}_2X_i)^2$$

$$\partial S/\partial\hat{\beta}_1=\sum[2\hat{\beta}_1-2(Y_i-\hat{\beta}_2X_i)]=0$$

$$\partial S/\partial\hat{\beta}_2=\sum[2X_i^2\hat{\beta}_2-2(Y_i-\hat{\beta}_1)X_i]=0$$

$$\sum Y_i-n\hat{\beta}_1-(\sum X_i)\hat{\beta}_2=0$$

$$\sum X_iY_i-(\sum X_i)\hat{\beta}_1-(\sum X_i^2)\hat{\beta}_2=0$$

$$\sum Y_i=n\hat{\beta}_1+\left(\sum X_i\right)\hat{\beta}_2 \quad \text{（式3.1.4）}$$

$$\sum X_iY_i=\left(\sum X_i\right)\hat{\beta}_1+\left(\sum X_i^2\right)\hat{\beta}_2 \quad \text{（式3.1.5）}$$

*正规方程组

用OLS方法估计总体参数β_i
——得到估计值得两种计算形式

（式3.1.6）

（式3.1.7）

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

（公式1：FF形式）
（Favorite Five）

$$\begin{cases}\sum x_iy_i=\sum(X_i-\bar{X})(Y_i-\bar{Y})=\sum X_iY_i-\dfrac{1}{n}\sum X_i\sum Y_i\\ \sum x_i^2=\sum(X_i-\bar{X}_i)^2=\sum X_i^2-1/n\left(\sum X_i\right)^2\end{cases}$$

（式3.1.6）

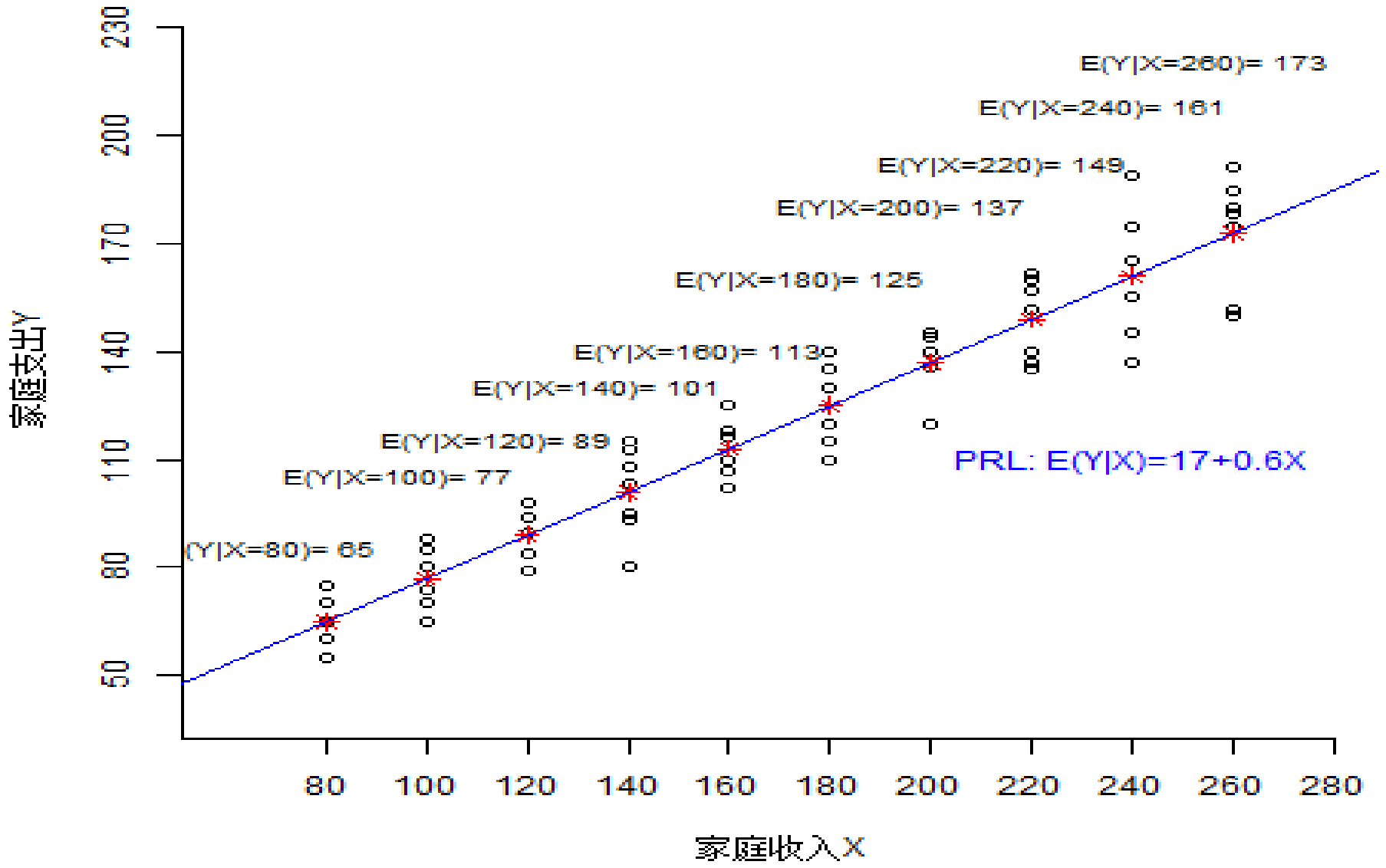
（式3.1.7）

$$\begin{cases}\hat{\beta}_2=\dfrac{\sum x_iy_i}{\sum x_i^2}\\ \hat{\beta}_1=\bar{Y}_i-\hat{\beta}_2\bar{X}_i\end{cases}$$

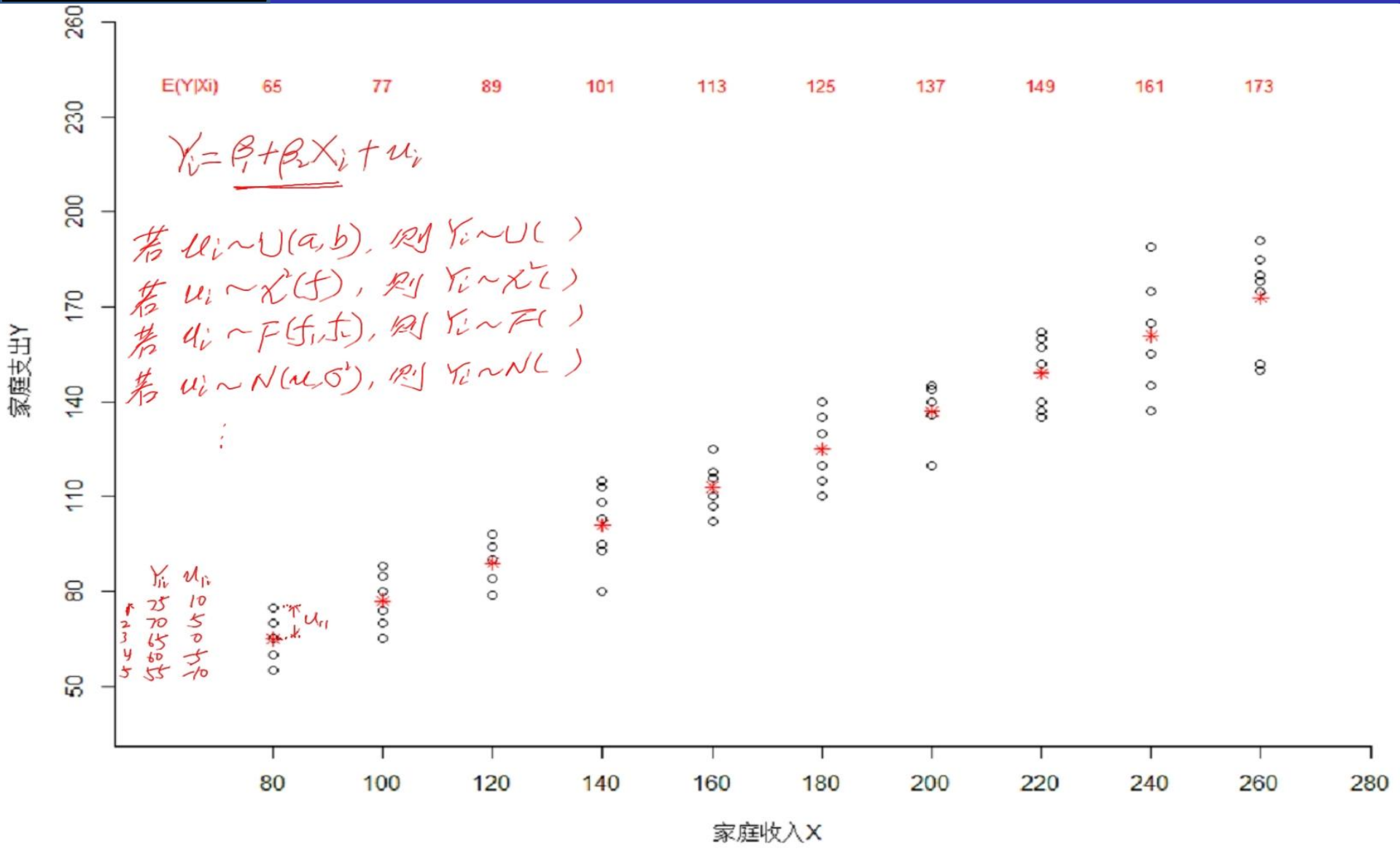
$$x_i=X_i-\bar{X}$$
$$y_i=Y_i-\bar{Y}$$

（公式2：离差形式）

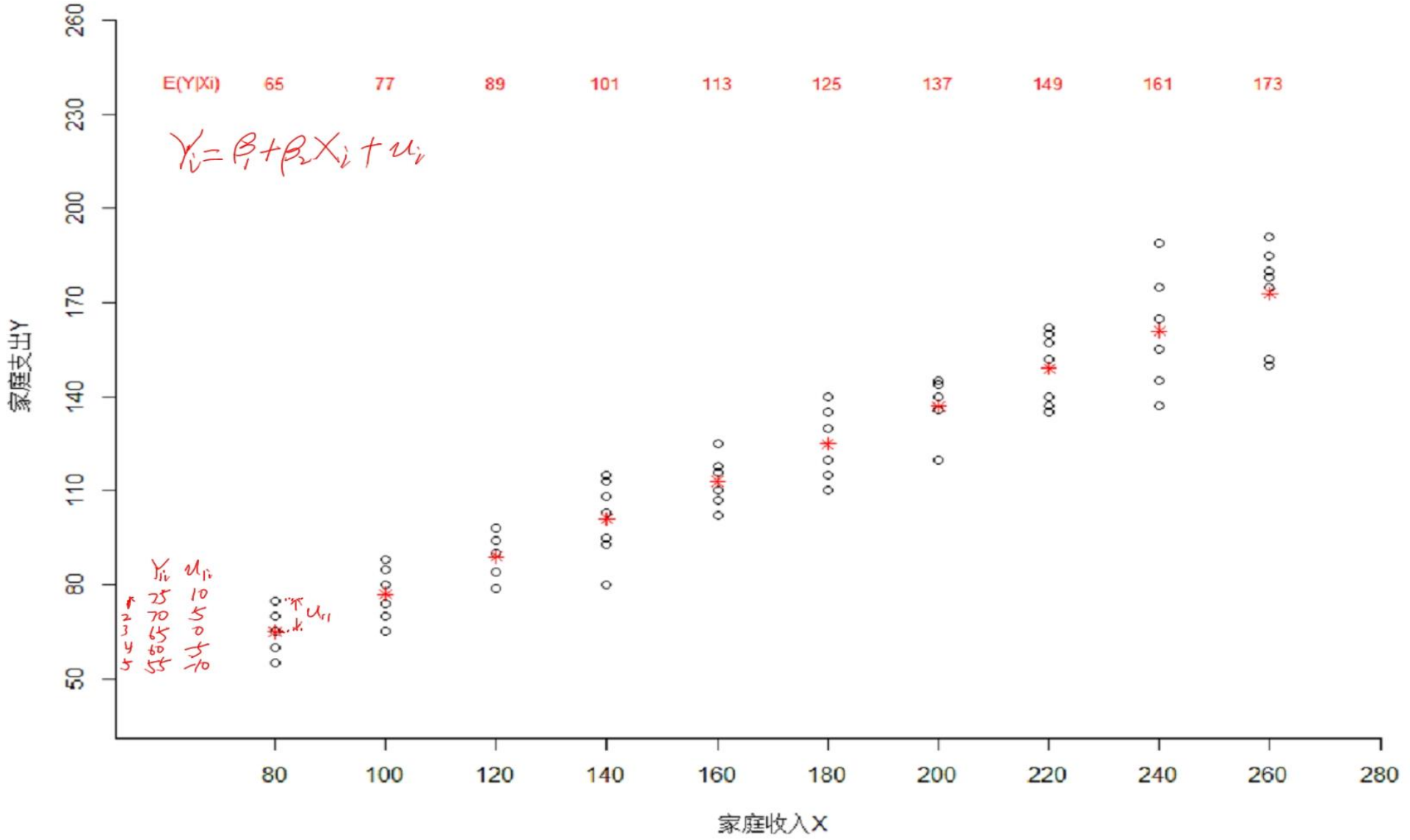
不得不说的故事：关于Y_i和μ_i的秘密 1 *



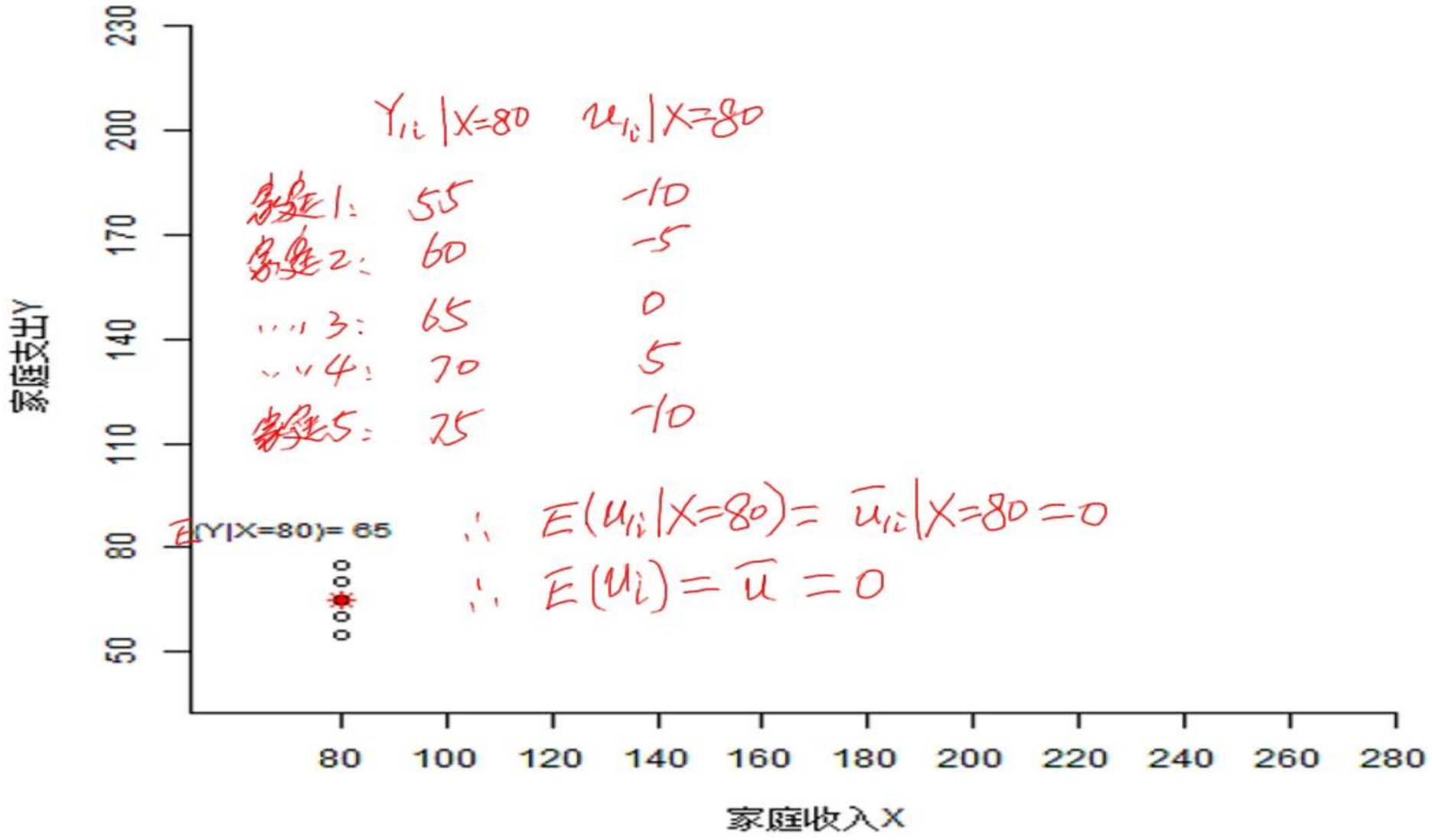
不得不说的故事：关于Y_i和μ_i的秘密 2 *



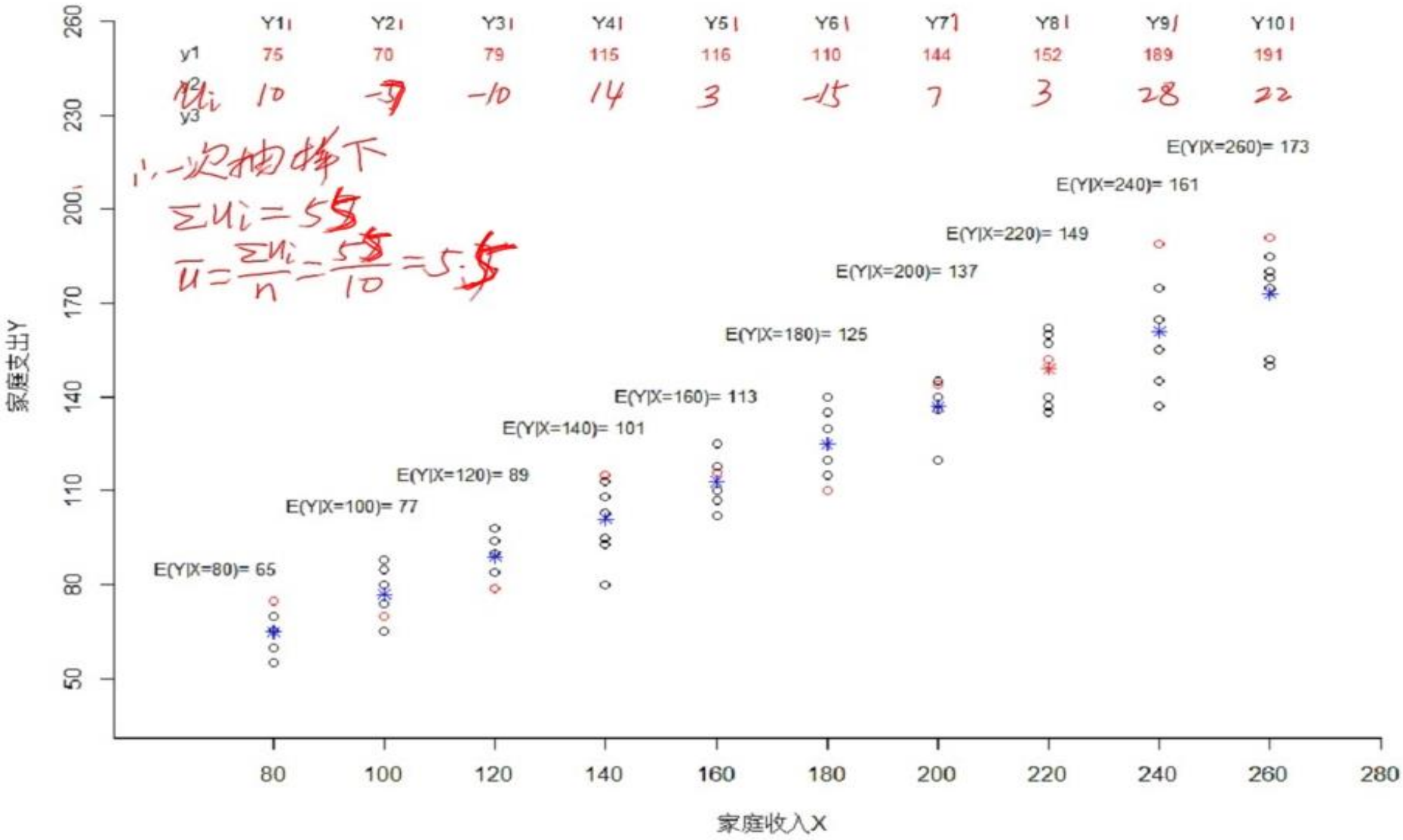
不得不说的故事：关于Y_i和μ_i的秘密 3 *



不得不说的故事：关于Y_i和μ_i的秘密 4 *



不得不说的故事：关于Y_i和μ_i的秘密 5 *



用OLS方法估计总体参数σ^2
——简要过程(要求记住结果！)

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

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

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

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

← $Y_i = \beta_1 + \beta_2 X_i + u_i$

$\bar{Y} = \beta_1 + \beta_2 \bar{X} + \bar{u}$

$y_i = \beta_2 x_i + (u_i - \bar{u})$

（思考）
此处μ_i为特指
此处$\bar{\mu}$不为零
Why？？

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

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

$$= A + B + C$$

$$= \sigma^2 + (n-1)\sigma^2 - 2\sigma^2$$

$$= (n-2)\sigma^2$$

← 具体推导见下一页附录

（式3.3.8）

（思考）

回归误差方差 $\hat{\sigma}^2 = \dfrac{\sum e_i^2}{(n-2)}$

回归误差标准差se $\hat{\sigma} = \sqrt{\dfrac{\sum e_i^2}{(n-2)}}$

用OLS方法估计总体参数σ^2
——附录（可自学）*

$$A=\sum x_i^2 E(\hat{\beta}_2-\beta_2)^2=\sum x_i^2 var(\hat{\beta}_2)=\sigma^2$$

$$\begin{aligned}B&=E[\sum(u_i-\bar{u})^2]\\&=E(\sum u_i^2)-2E(\sum u_i\bar{u})+nE(\bar{u}^2)\\&=n\sigma^2-2E[\sum(u_i\frac{\sum u_i}{n})]+nE(\frac{\sum u_i}{n})^2\\&=n\sigma^2-2E(\sum\frac{u_i^2}{n})+nE\frac{u_1^2+u_2^2+\cdots+u_n^2}{n^2}\\&=n\sigma^2-2\sigma^2+\sigma^2=(n-1)\sigma^2\end{aligned}$$

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

$$\begin{aligned}\hat{\beta}_2&=\sum k_i\cdot Y_i\\&=\sum k_i(\beta_1+\beta_2X_i+u_i)\\&=\beta_1\sum k_i+\beta_2\sum k_iX_i+\sum k_iu_i\\&=\beta_2+\sum k_iu_i\end{aligned}$$

理解：OLS方法下的“估计值”与“估计量”

（式3.1.6）

（式3.1.7）

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

（形式1：FF形式）
（Favorite Five）

$$\begin{cases}\hat{\beta}_2=\dfrac{\sum x_iy_i}{\sum x_i^2}\\ \hat{\beta}_1=\bar{Y}_i-\hat{\beta}_2\bar{X}_i\end{cases}$$

（形式2：离差形式）

- 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**”。

讨论1：用OLS方法得到参数估计量的特征

- 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。[我们以后还将考虑区间估计量(interval Estimators)]
- 一旦从样本数据得到OLS估计值，便容易画出样本回归线。

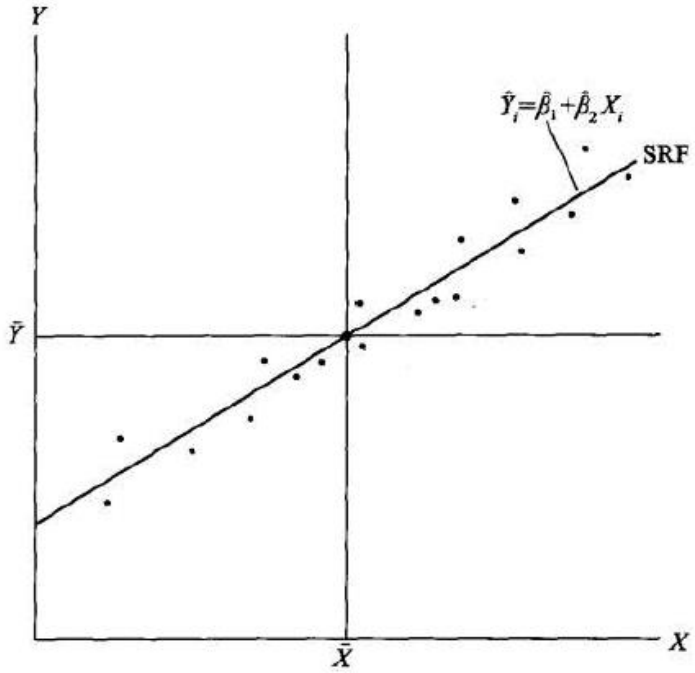


Figure 3-2

讨论2：用OLS方法得到的SRF的特征(1/3)

- A. 它穿过Y和X的样本均值点：

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

[观察图形]

- B. Y的估计值(=$\hat{Y}_i$)的均值等于实际Y值的均值：

（式3.1.9）

$$\hat{Y}_i=\hat{\beta}_1+\hat{\beta}_2X_i=(\bar{Y}-\hat{\beta}_2\bar{X})+\hat{\beta}_2X_i=\bar{Y}+\hat{\beta}_2(X_i-\bar{X})$$

[$\hat{\beta}_1$公式]

[SRF]

- 两边分别处理 $1/n\sum$ ，得到：

小技巧！

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

即： $\overline{\hat{Y}_i}=\bar{Y}_i$

[代数运算]

讨论2：用OLS方法得到的SRF的特征(2/3)

- C1. 残差(=e_i)的均值为零：

$$\sum[2\hat{\beta}_1 - 2(Y_i - \hat{\beta}_2 X_i)] = 0$$

[正规方程1]

$$\sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$$

$$\sum(Y_i - \hat{Y}_i) = 0$$

[SRF定义]

$$\sum e_i = 0$$

[残差定义]

- C2. SRF也能写成离差形式：

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

[SRM定义]

（式2.6.2）

$$\therefore \frac{1}{n}\sum Y_i = \frac{1}{n}\sum(\hat{\beta}_1 + \hat{\beta}_2 X_i + e_i)$$

[代数运算]

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

[SRF回归线]

（式3.1.12）

$$\therefore Y_i - \bar{Y} = (\hat{\beta}_1 + \hat{\beta}_2 X_i + e_i) - (\hat{\beta}_1 + \hat{\beta}_2 \bar{X}) = \hat{\beta}_2(X_i - \bar{X}) + e_i$$

[(式2.6.2)-(式3.1.12)]

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

SRM离差形式

约定(式3.1.13)

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

SRF离差形式

约定(式3.1.14)

讨论2：用OLS方法得到的SRF的特征(3/3)

- D. 残差(e_i)和Y的预测值($\hat{Y}_i$)不相关。
  - 利用离差形式SRF来进行证明：

（式3.1.15）

$$\begin{aligned}\sum \hat{y}_i \cdot e_i &= \sum \hat{\beta}_2 x_i \cdot e_i \\ &= \hat{\beta}_2 \sum x_i \cdot (y_i - \hat{\beta}_2 x_i) \\ &= \hat{\beta}_2 \sum (x_i \cdot y_i - \hat{\beta}_2 x_i^2) \\ &= \hat{\beta}_2^2 \sum x_i^2 - \hat{\beta}_2^2 \sum x_i^2 \\ &= 0\end{aligned}$$

[SRF离差]

[SRF-s离差]

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

[代数运算]

- E. 残差(e_i)和X_i不相关

离差公式
——一个总结性概览

离差定义与符号

$$x_i = X_i - \bar{X};\quad y_i = Y_i - \bar{Y};\quad \hat{y}_i = \hat{Y}_i - \bar{\hat{Y}} = \hat{Y}_i - \bar{Y}$$

PRM及其离差形式

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

SRM及其离差形式

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

SRF及其离差形式

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

残差的离差形式

$$e_i = Y_i - \hat{Y} = y_i - \hat{\beta}_2 x_i$$

§ 2.3.2 经典线性回归模型(CLRM)

- 经典（又称高斯或标准）线性回归模型(classical linear regression model, CLRM)
  - 又称高斯或标准线性回归模型
  - 成为大部分计量经济学理论的基石，有7个基本假设。
  - 本章以双变量回归模型为讨论基础：

（式3.1.15）

- 假设1：线性回归假设
  - 模型中参数必须线性，变量可以不是线性；

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

经典线性回归模型(CLRM)：基本假设(2)

- 假设2：X是固定的**或**独立于误差项 [本科生学习]
  - 固定回归元：X取值是固定的，但随机取Y
  - ~~随机回归元：X与Y同时随机取值，要求X独立于误差项~~ [研究生学习]

$$\text{cov}(X_i, \mu_i) = 0, i = 1, 2, \cdots, n$$
$$E(X_i \mu_i) = 0, i = 1, 2, \cdots, n$$

Table 2-4/5

随机样本1	X	80	100	120	140	160	180	200	220	240	260
	Y	70	65	90	95	110	115	120	140	155	150
随机样本2	X	80	100	120	140	160	180	200	220	240	260
	Y	55	88	90	80	118	120	145	135	145	175

随机样本3	X	80	80	100	120	140	160	200	220	240	260
	Y	55	75	74	84	103	116	136	140	155	150
随机样本4	X	80	80	100	140	160	180	200	200	260	260
	Y	65	70	74	80	125	120	136	144	152	175

思考：

为什么先学习情形1？X值固定不变现实么？
OLS下两种情形得到的结果相同么？

经典线性回归模型(CLRM)：基本假设(3)

- 假设3:干扰项μ_i的均值为零
  - 若是固定回归元：μ_i的条件期望为零

$$E(\mu_i|X_i) = 0$$

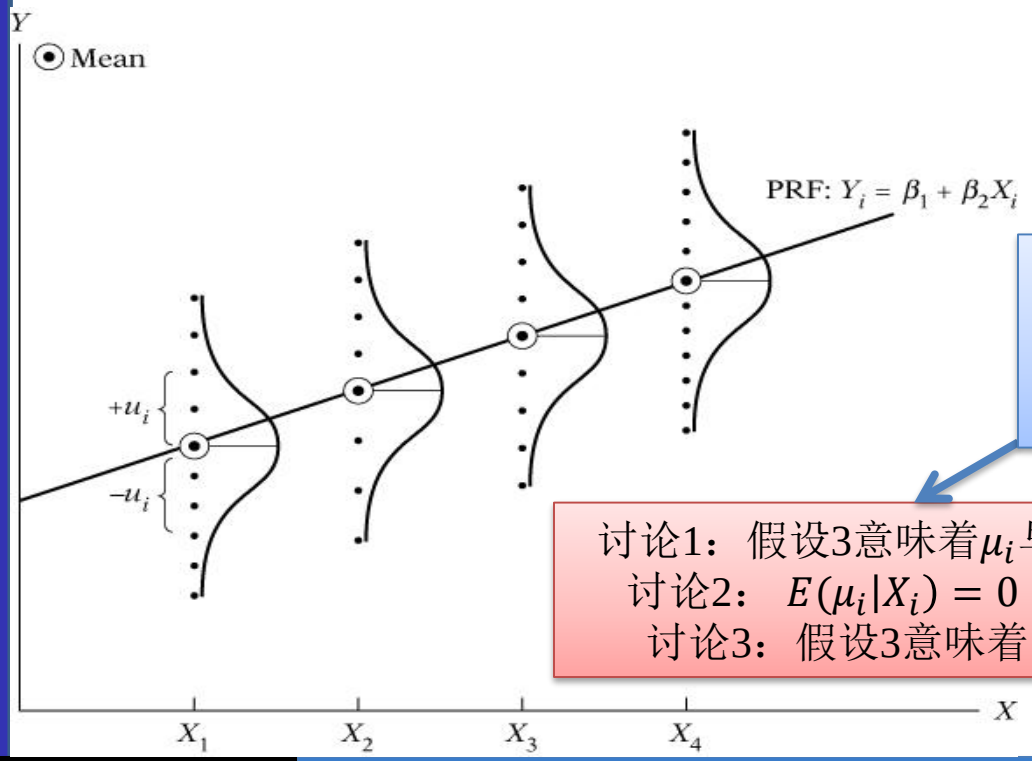
[本科生学习]

  - ~~若是随机回归元：μ_i的无条件期望为零~~

$$E(\mu_i) = 0$$

[研究生学习]

Figure 3-3



如果μ_i与X_i 不相关，则
$E(\mu_i|X_i)$也可能=0；
但如果μ_i与X_i 相关，则
$E(\mu_i|X_i) = 0$，即违背假设3

讨论*：

讨论1：假设3意味着μ_i与X_i 不相关。反之却不成立
讨论2：$E(\mu_i|X_i) = 0 \Leftrightarrow E(Y_i|X_i) = \beta_1 + \beta_2 X_2$
讨论3：假设3意味着“模型是正确设定的”

经典线性回归模型(CLRM)：基本假设(4-1)

- 假设4:同方差性或μ_i的方差相等
  - 给定X值，对所有的观测，μ_i的方差都是相同的

$$\begin{aligned}\text{var}(u_i \mid X_i) &= E\{[u_i - E(u_i) \mid X_i]\}^2 \\ &= E(u_i^2 \mid X_i) \\ &= E(u_i^2) \\ &= \sigma^2\end{aligned}$$

[方差定义]

[由于假设3]

[由于假设2]

Figure 3-4

- 同方差性(homoscedasticity)：

$$\text{var}(u_i \mid X_i) = \sigma^2$$

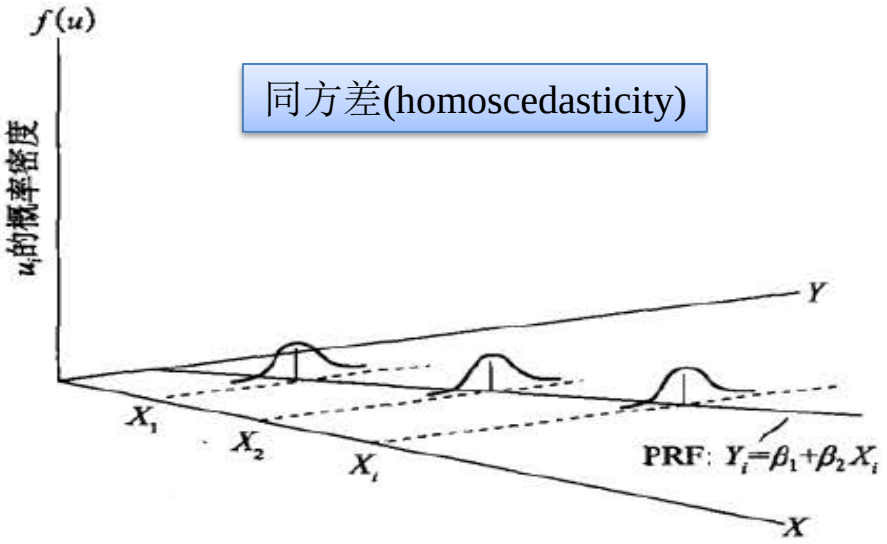
讨论*：

- 异方差性(heteroscedasticity)：

$$\text{var}(u_i \mid X_i) = \sigma_i^2$$

经典线性回归模型(CLRM)：基本假设(4-2)

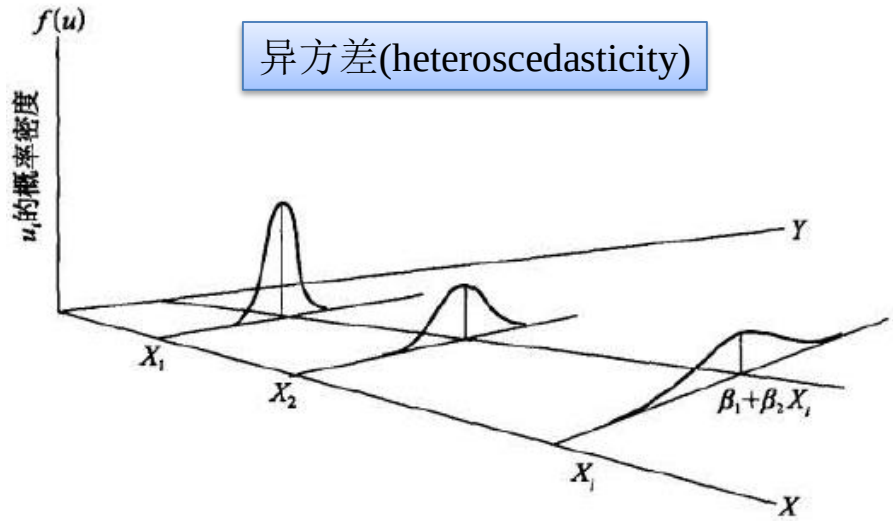
Figure 3-4



同方差(homoscedasticity)

讨论1：不同X 值的Y 总体均有同样的方差。

Figure 3-5



异方差(heteroscedasticity)

讨论*：

讨论1: $\text{var}(\mu_1|X_1) < \text{var}(\mu_2|X_2)$ 意味着来自$X = X_1$ 的Y 总体，相比来自$X = X_2$ 的Y 总体，更靠近PRF。

讨论2：我们要认为那些**离均值较近的Y总体**的**样本**比远为分散的Y总体的**样本**更为可取吗?

讨论3：如果出现异方差，会对OLS估计产生什么后果？

经典线性回归模型(CLRM)：基本假设(5-1)

- 假设5：各个干扰之间无自相关 [本科生学习]
  - 给定任意两个X值：X_i和X_j，μ_i和μ_j的的相关系数为零。也就是说观测是相互独立的。

$$Cov(u_i,u_j)=E\{[u_i-E(u_i)]\cdot[u_j-E(u_j)]\}$$
$$=E(u_iu_j)=0 \quad \text{其中},(i\neq j)$$

[若X为固定元]

~~[若X为随机元]~~

$$Cov(u_i,u_j\mid X_i,X_j)=E\{[u_i-E(u_i)\mid X_i]\cdot[u_j-E(u_j)\mid X_j]\}$$
$$=E(u_i\mid X_i)(u_j\mid X_j)=0 \quad \text{其中},(i\neq j)$$

- 无序列相关(no serial correlation)：时间序列数据问题
- 截面数据不会出现。Why？

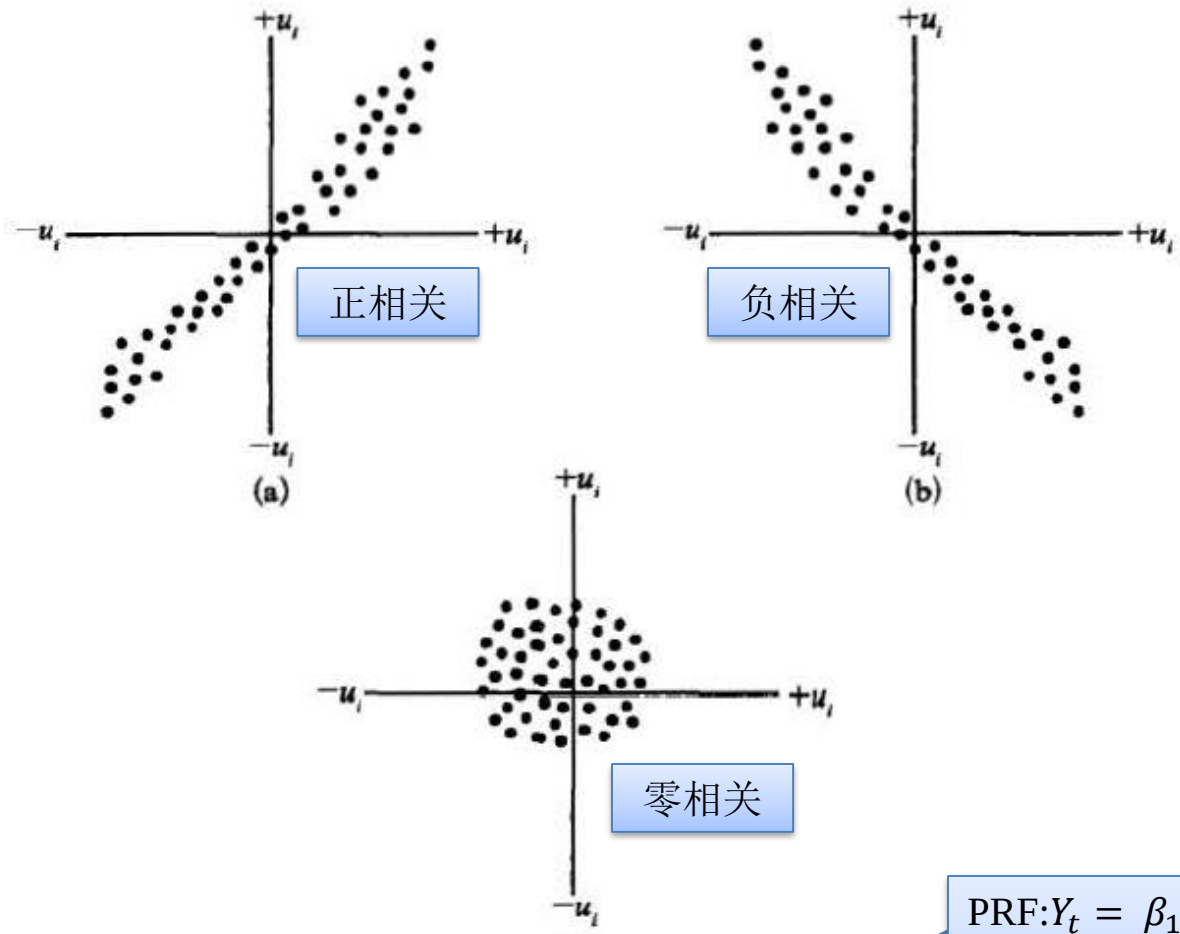
Table 2-4/5

思考：

为什么先学习情形1？X值固定不变现实么？
OLS下两种情形得到的结果相同么？

经典线性回归模型(CLRM)：基本假设(5-2)

Figure 3-6



思考：

问1：假设5的目的和用处是什么？
问2：如果出现自相关，会对OLS估计产生什么影响？

PRF:$Y_t = \beta_1 + \beta_2 X_t + \mu_t$
如果，μ_t与μ_{t-1}正相关，
那么，Y_t不仅依赖于X_t，还依赖于μ_{t-1}

经典线性回归模型(CLRM)：基本假设(6、7)

- 假设6:观测次数n，要大于待估计参数个数
  - 否则方程无法解出，参数不能估计出来

- 假设7:X变量的性质假设
  - Var(X)是有限的正数;
    - 例如X取值不能全部相同(此时Var(X)=0)
    - 估计值公式中分母为0，无法求解!
  - X变量没有异常值(outlier)，即没有一个X值对于其他值过大或过小

Figure 3-6

思考:

经典线性回归模型(CLRM)：基本假设（大讨论）

思考：

- 所有这些假设有多真实?
  - “假定无关紧要论”——弗里德曼
- 上述说有假设都是针对PRF，而不是SRF！
  - 例如：PRF中假设5——$\text{cov}(\mu_i, \mu_j) = 0$，但在SRF中，可能就会出现$\text{cov}(e_i, e_j) \neq 0$
- 前面提到的OLS方法正是试图“复制”CLRM的假设！
  - OLS方法中$\sum e_i = 0$，(即$\bar{e}_i$=0)就类似于CLRM的假设3——$E(\mu_i|X_i) = 0$
  - OLS方法中$\sum e_i X_i$=0，就类似于CLRM的假设2——$\text{cov}(\mu_i, X_i) = 0$

OLS估计量的精度
——[$\hat{\beta}_2$]的方差和样本方差

- $\hat{\beta}_2$的方差和标准差：

（思考）

$$\text{var}(\hat{\beta}_2) = \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2}{\sum x_i^2} \quad （式3.3.1）$$

$$s.e.(\hat{\beta}_2) = \sigma_{\hat{\beta}_2} = \frac{\sigma}{\sqrt{\sum x_i^2}} \quad （式3.3.2）$$

- $\hat{\beta}_2$的样本方差和样本标准误：

$$S^2_{\hat{\beta}_2} = \frac{\hat{\sigma}^2}{\sum x_i^2} \quad （式3.3.3）$$

$$S_{\hat{\beta}_2} = \frac{\hat{\sigma}}{\sqrt{\sum x_i^2}} \quad （式3.3.4）$$

回归误差方差 $\hat{\sigma}^2 = \frac{\sum e_i^2}{(n-2)}$

回归误差标准se $\hat{\sigma} = \sqrt{\frac{\sum e_i^2}{(n-2)}}$

OLS估计量的精度
——[$\hat{\beta}_1$]的方差和样本方差

（思考）

- $\hat{\beta}_1$的方差和标准差：

（式3.3.1）

$$\text{var}(\hat{\beta}_1) = \sigma^2_{\hat{\beta}_1} = \frac{\sum X_i^2}{n\sum x_i^2}\sigma^2$$

（式3.3.2）

$$s.e.(\hat{\beta}_1) = \sigma_{\hat{\beta}_1} = \sqrt{\frac{\sum X_i^2}{n\sum x_i^2}} \cdot \sigma$$

- $\hat{\beta}_1$的样本方差和样本标准差：

（式3.3.3）

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

（式3.3.4）

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

回归误差方差 $\hat{\sigma}^2 = \frac{\sum e_i^2}{(n-2)}$

回归误差标准se $\hat{\sigma} = \sqrt{\frac{\sum e_i^2}{(n-2)}}$

最小二乘法（OLS）的估计量：——$\hat{\beta}_2$方差[var($\hat{\beta}_2$)]推导过程1/2

（思考）

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

$$= \frac{\sum x_i \left(Y_i - \bar{Y}\right)}{\sum x_i^2}$$

$$= \frac{\sum x_i Y_i - \sum x_i \bar{Y}}{\sum x_i^2}$$

$$= \frac{\sum x_i Y_i - \bar{Y}\sum x_i}{\sum x_i^2} \quad \leftarrow \sum x_i = \sum (X_i - \bar{X}) = 0$$

$$= \sum \left(\frac{x_i}{\sum x_i^2} \cdot Y_i \right) \quad \leftarrow 令 k_i = \frac{x_i}{\sum x_i^2}$$

$$= \sum k_i Y_i$$

最小二乘法（OLS）的估计量：——$\hat{\beta}_2$方差[$\mathrm{var}(\hat{\beta}_2)$]推导过程2/2

$$\mathrm{var}(\hat{\beta}_2) = \mathrm{var}\left(\sum\left(k_i \cdot Y_i\right)\right)$$

$$= \sum\left(k_i^2 \cdot \mathrm{var}(Y_i)\right)$$

$$= \sum\left(k_i^2 \cdot \mathrm{var}(\beta_1 + \beta_2 X_i + \mu_i)\right)$$

$$= \sum\left(k_i^2 \cdot \mathrm{var}(\mu_i)\right)$$

$$= \sum\left(\left(\frac{x_i}{\sum x_i^2}\right)^2 \cdot \sigma^2\right)$$

$$= \frac{\sigma^2}{\sum x_i^2}$$

$k_i = \dfrac{x_i}{\sum x_i^2}$

[方差性质]

$k_i = \dfrac{x_i}{\sum x_i^2}$

最小二乘法（OLS）的估计量：——$\hat{\beta}_1$方差[var($\hat{\beta}_1$)]推导过程1/2

（思考）

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

$$= \frac{1}{n}\sum Y_i - \sum\left[(k_i Y_i)\cdot \bar{X}\right]$$

$$= \sum\left[\left(\frac{1}{n} - \bar{X}k_i\right)\cdot Y_i\right]$$

$$= \sum w_i \cdot Y_i$$

$\hat{\beta}_2 = \sum k_i Y_i$

$k_i = \dfrac{x_i}{\sum x_i^2}$

令$w_i = \dfrac{1}{n} - \bar{X}k_i$

最小二乘法（OLS）的估计量：——$\hat{\beta}_1$方差[var($\hat{\beta}_1$)]推导过程2/2

$$\mathrm{var}(\hat{\beta}_1)=\mathrm{var}\left(\sum w_iY_i\right)=\sum\left(w_i^2\cdot\mathrm{var}(Y_i)\right)$$

[方差性质] →

$$=\sum\left(w_i^2\cdot\mathrm{var}(\beta_1+\beta_2X_i+\mu_i)\right)=\sum\left(w_i^2\cdot\mathrm{var}(\mu_i)\right)$$

$$w_i=\frac{1}{n}-\overline{X}k_i$$

$$=\sum\left(\left(\frac{1}{n}-\overline{X}k_i\right)^2\cdot\sigma^2\right)=\sigma^2\sum\left(\frac{1}{n^2}-\frac{2\overline{X}k_i}{n}+\overline{X}^2k_i^2\right)$$

$$\sum k_i=\sum\left(\frac{x_i}{\sum x_i^2}\right)=\frac{\sum x_i}{\sum x_i^2}=0$$

$$=\sigma^2\left(\frac{1}{n}-\frac{2\overline{X}\sum k_i}{n}+\overline{X}^2\sum k_i^2\right)=\sigma^2\left(\frac{1}{n}+\overline{X}^2\sum\left(\frac{x_i}{\sum x_i^2}\right)^2\right)$$

$$k_i=\frac{x_i}{\sum x_i^2}$$

[代数运算] →

$$=\sigma^2\left(\frac{1}{n}+\overline{X}^2\cdot\frac{\sum x_i^2}{\left(\sum x_i^2\right)^2}\right)=\sigma^2\left(\frac{1}{n}+\frac{\overline{X}^2}{\sum x_i^2}\right)$$

$$\sum x_i^2+n\overline{X}^2$$
$$=\sum\left(X_i-\overline{X}\right)^2+n\overline{X}^2$$
$$=\left(\sum X_i^2-2\overline{X}\sum X_i+n\overline{X}^2\right)+n\overline{X}^2$$
$$=\sum X_i^2$$

$$=\frac{\sum x_i^2+n\overline{X}^2}{n\sum x_i^2}\cdot\sigma^2=\frac{\sum X_i^2}{n\sum x_i^2}\cdot\sigma^2$$

§ 2.3.4 最小二乘估计的性质：BLUE

- **最优线性无偏估计量(BLUE)需要满足的条件：**
  - 线性的(linear)：估计量是因变量Y的线性函数。
  - 无偏的(unbiased)：估计量的均值或期望值$E(\hat{\beta}_i)$ 等于真值β_i。
  - 有效的(efficient)：是所有线性无偏估计量中有最小方差的那个估计量

（定理）

- **高斯-马尔可夫定理(Gauss-Markov Theorem)：**
  - 在给定经典线性回归模型**(CLRM)**的假定下，最小二乘**(OLS)**估计量，在无偏线性估计量一类中，有最小方差，就是说它们是一个最优线性无偏估计量**(BLUE)**

（思考）

问1：为什么最小二乘法（OLS）被计量学家奉为神明？还有其他选择吗？
问2：OLS得到的BLUE为到底有什么值得你称赞？
问3：OLS得到BLUE还需要CLRM假设以外的更多假设吗？(正态性？？)

最小二乘法（OLS）的性质：
——高斯-马尔可夫定理(Gauss-Markov Theorem)：图解

Figure 3-7

(a)$\hat{\beta}_2$是β_2的一个线性无偏估计量

(b) β_2^* 也是β_2的一个线性无偏估计量

(c)那么哪一个估计量更能为我们接受呢？

（思考）

问1：什么是抽样分布？
问2：怎样获得估计量分布？
问3：没有比OLS估计量更好的估计量了吗？

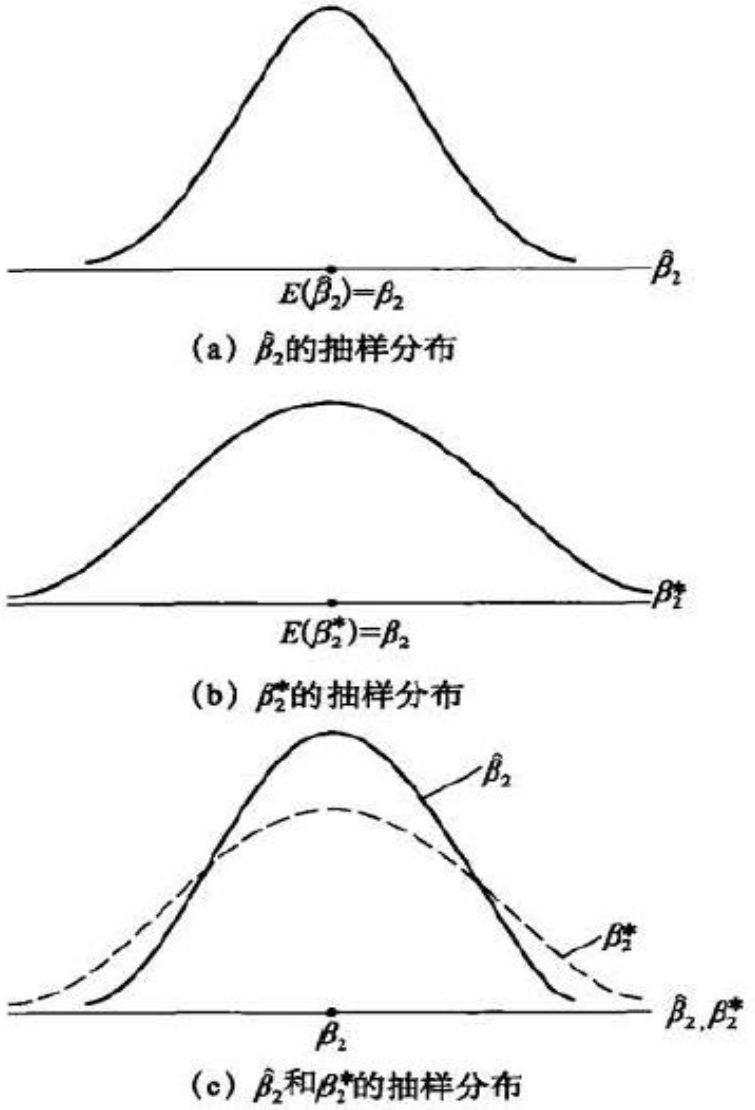


图3-7 OLS方法下两个估计量的抽样分布

1.线性性（Linearity）

——$\hat{\beta}_1$和$\hat{\beta}_2$对Y_i是线性的

$$\hat{\beta}_2 = \sum k_i Y_i \qquad k_i = \frac{x_i}{\sum x_i^2}$$

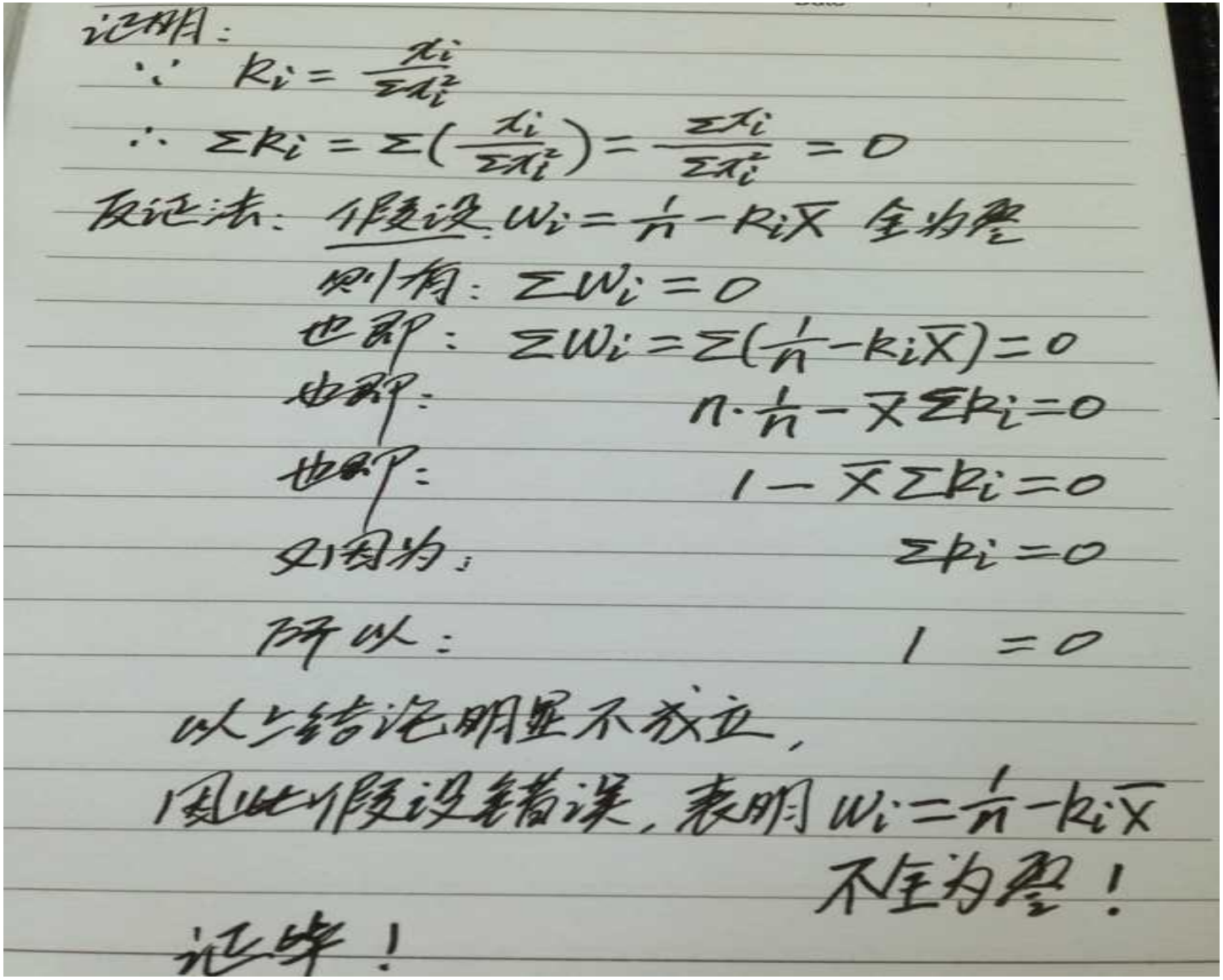
$\because k_i = \dfrac{x_i}{\sum x_i^2}$ is not all equal to 0

$\therefore \hat{\beta}_2$ is linear to Y_i

$$\hat{\beta}_1 = \sum w_i \cdot Y_i \qquad w_i = \frac{1}{n} - \overline{X} k_i$$

$\because w_i = \dfrac{1}{n} - \overline{X} k_i$ is not all equal to 0 (why??)

$\therefore \hat{\beta}_1$ is linear to Y_i


证明：
$\because k_i = \frac{x_i}{\sum x_i^2}$
$\therefore \sum k_i = \sum\left(\frac{x_i}{\sum x_i^2}\right) = \frac{\sum x_i}{\sum x_i^2} = 0$
反证法：假设 $w_i = \frac{1}{n} - k_i\bar{X}$ 全为零
则有：$\sum w_i = 0$
也即：$\sum w_i = \sum\left(\frac{1}{n} - k_i\bar{X}\right) = 0$
也即：$n \cdot \frac{1}{n} - \bar{X}\sum k_i = 0$
也即：$1 - \bar{X}\sum k_i = 0$
又因为：$\sum k_i = 0$
所以：$1 = 0$
以上结论明显不成立，
因此假设错误，表明 $w_i = \frac{1}{n} - k_i\bar{X}$
不全为零！
证毕！

2. 无偏性

$$\hat{\beta}_2 = \sum k_i Y_i$$

$$k_i = \frac{x_i}{\sum x_i^2}$$

$$= \sum k_i (\beta_1 + \beta_2 X_i + \mu_i)$$

$$= \beta_1 \sum k_i + \beta_2 \sum k_i X_i + \sum k_i \mu_i$$

$$= 0 + \beta_2 + \sum k_i \mu_i$$

$$\sum k_i = \sum \left(\frac{x_i}{\sum x_i^2} \right) = \frac{\sum x_i}{\sum x_i^2} = 0$$

$$\mathrm{E}(\hat{\beta}_2) = \beta_2 + E(\sum k_i \mu_i)$$

$$= \beta_2 + \sum [k_i E(\mu_i)]$$

$$= \beta_2$$

$$\sum k_i X_i = \sum \left[\frac{x_i}{\sum x_i^2} \cdot X_i \right]$$

$$= \frac{\sum (x_i X_i)}{\sum x_i^2}$$

$$= \frac{\sum \left[x_i \left(x_i + \overline{X} \right) \right]}{\sum x_i^2}$$

$$= \frac{\sum \left[x_i^2 + x_i \overline{X} \right]}{\sum x_i^2}$$

$$= \frac{\sum x_i^2 + \overline{X} \sum x_i}{\sum x_i^2} = 1$$

2. 无偏性（Unbiasedness）
——$\hat{\beta}_1$是无偏的

$$\hat{\beta}_1 = \sum w_i \cdot Y_i$$
$$= \sum w_i \cdot (\beta_1 + \beta_2 X_i + \mu_i)$$
$$= \beta_1 \sum w_i + \beta_2 \sum w_i X_i + \sum w_i \mu_i$$
$$= \beta_1 + 0 + \sum \mu_i w_i$$

$$w_i = \frac{1}{n} - \bar{X} k_i \qquad k_i = \frac{x_i}{\sum x_i^2}$$

$$\sum w_i = \sum \left(\frac{1}{n} - \bar{X} k_i\right) = 1 - \bar{X} \sum k_i = 1$$

$$\sum w_i X_i = \sum \left[\left(\frac{1}{n} - \bar{X} k_i\right) X_i\right]$$
$$= \sum \left[\frac{X_i}{n} - \bar{X} k_i X_i\right]$$
$$= \bar{X} - \bar{X} \cdot \sum (k_i X_i)$$
$$= \bar{X} - \bar{X}$$
$$= 0$$

$$E(\hat{\beta}_1) = E\left(\beta_1 + \sum w_i \mu_i\right)$$
$$= \beta_1 + \sum [w_i E(\mu_i)]$$
$$= \beta_1$$

3. 最优性（Best）
——$\hat{\beta}_1$和$\hat{\beta}_2$的方差是最小的

- 证明：已知方差

$$\text{方差}\ \text{var}(\hat{\beta}_2)=\sigma^2_{\hat{\beta}_2}=\frac{\sigma^2}{\sum x_i^2}$$

$$\text{方差}\ \text{var}(\hat{\beta}_1)=\sigma^2_{\hat{\beta}_1}=\frac{\sum X_i^2}{n\sum x_i^2}\sigma^2$$

假设存在用其他方法估计的线性无偏估计量：

$$\hat{\beta}_2^*=\sum\left[(k_i+d_i)Y_i\right]=\sum c_iY_i$$

$$\hat{\beta}_1^*=\sum\left[(w_i+g_i)Y_i\right]=\sum h_iY_i$$

其中，di和gi为不全为零的常数，则可以证明（此处略）：

$$\text{var}(\hat{\beta}_2^*)\geq\text{var}(\hat{\beta}_2)$$

$$\text{var}(\hat{\beta}_1^*)\geq\text{var}(\hat{\beta}_1)$$

因此，方差最小性得以证明！

§ 2.3.5 判定系数

——判定系数(r^2 或R^2)：直观印象

- 拟合优度(goodness of fit)：
  - 样本回归线对一组数据拟合得有多好。
- 判定系数(coefficient of determination)：
  - 双变量回归(一元回归)：r^2
  - 多变量回归(多元回归)：R^2

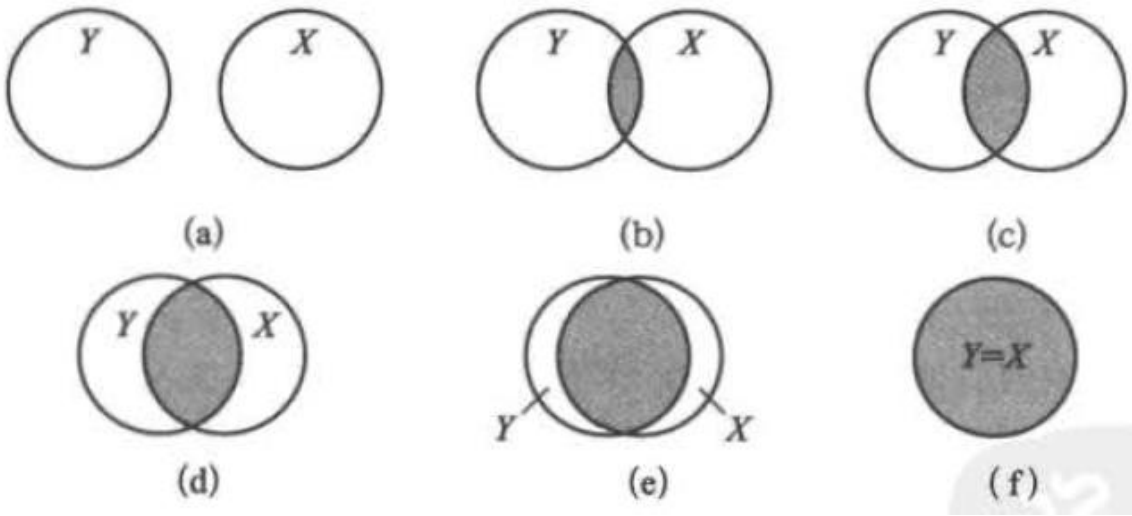


图3-8 变量的变异与r^2

Figure 3-8

（思考）

问1：采用普通最小二乘估计方法，已经保证了模型最好地拟合了样本观测值，为什么还要计算拟合优度(R^2)?

拟合优度(goodness of fit)的度量：
——判定系数(r^2 或R^2)：方差分解

Figure 3-8

① Y_i变异的分解
② 平方和分解
③ 方差分解

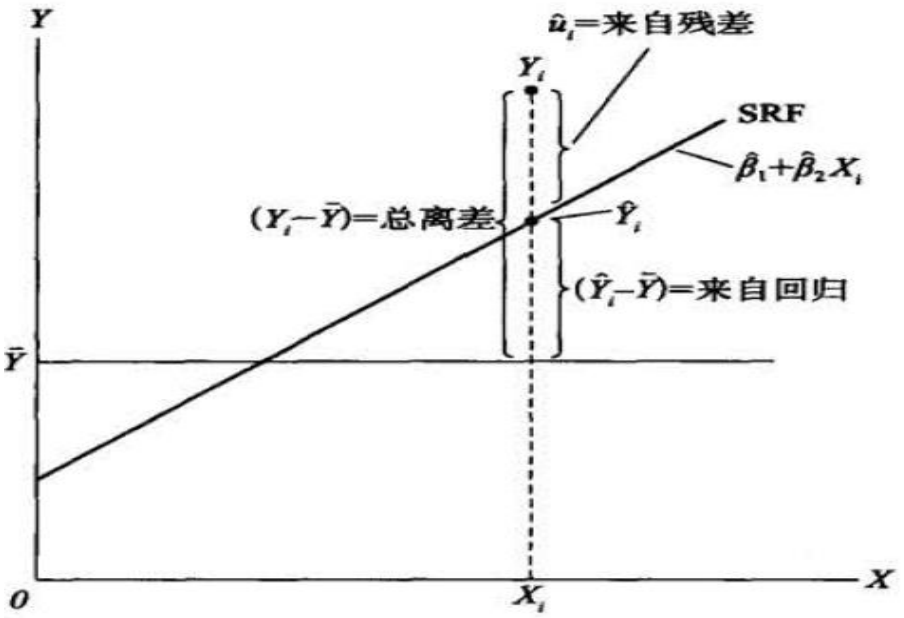


图3-9 y_i 变异的分解

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

$$\sum y_i^2=\sum \hat{y}_i^2+\sum e_i^2+2\sum \hat{y}_i e_i$$
$$=\sum \hat{y}_i^2+\sum e_i^2$$

$$\sum \hat{y}_i e_i=0$$

$$=\hat{\beta}_2^2\sum x_i^2+\sum e_i^2$$

（思考）

TSS = ESS + RSS

总离差平方和 = 回归平方和 + 剩余平方和

拟合优度(goodness of fit)的度量：
——判定系数(r^2 或R^2)：3种计算公式

（式 3.5.8）
残差-离差形式

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

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

$$S_x^2 = 1/n \cdot \sum(X_i - \bar{X})^2$$

$$S_y^2 = 1/n \cdot \sum(Y_i - \bar{Y})^2$$

（式 3.5.6/7）
系数-离差形式

$$r^2 = \frac{ESS}{TSS} = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = \frac{\sum(\hat{\beta}_2 x_i)^2}{\sum y_i^2} = \hat{\beta}_2^2 \left(\frac{\sum x_i^2}{\sum y_i^2}\right) = \hat{\beta}_2^2 \cdot \frac{S_x^2}{S_y^2}$$

（式 3.5.8）
离差形式

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

（讨论）

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

拟合优度(goodness of fit)的度量：
——判定系数(r^2 或R^2)：与相关系数r的区别和联系

Figure 3-7

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

$$\rho = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\cdot\mathrm{var}(Y)}} = \frac{E\left[(X-EX)(Y-EY)\right]}{\sqrt{E\left[(X-EX)^2\right]\cdot E\left[(Y-EY)^2\right]}}$$

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

- 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)！

（思考）

§ 2.3.6 一个数值例子
——受教育程度(X)与时均工资(Y)

Table 3-2

$n=13$，$\bar{X}=12$，$\bar{Y}=8.674708$

	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	$\hat{Y}_i$	e_i	e_i^2
1	6	4.4567	26.7402	36	19.86217	-6	-4.21801	25.30805	36	4.4567	4.330127	0.126573	0.016021
2	7	5.77	40.39	49	33.2929	-5	-2.90471	14.52354	25	5.77	5.054224	0.715776	0.512335
3	8	5.9787	47.8296	64	35.74485	-4	-2.69601	10.78403	16	5.9787	5.778321	0.200379	0.040152
4	9	7.3317	65.9853	81	53.75382	-3	-1.34301	4.029023	9	7.3317	6.502418	0.829282	0.687709
5	10	7.3182	73.182	100	53.55605	-2	-1.35651	2.713015	4	7.3182	7.226514	0.091686	0.008406
6	11	6.5844	72.4284	121	43.35432	-1	-2.09031	2.090308	1	6.5844	7.950611	-1.36621	1.866532
7	12	7.8182	93.8184	144	61.12425	0	-0.85651	0	0	7.8182	8.674708	-0.85651	0.733605
8	13	7.8351	101.8563	169	61.38879	1	-0.83961	-0.83961	1	7.8351	9.398804	-1.5637	2.445171
9	14	11.0223	154.3122	196	121.4911	2	2.347592	4.695185	4	11.0223	10.1229	0.899399	0.808918
10	15	10.6738	160.107	225	113.93	3	1.999092	5.997277	9	10.6738	10.847	-0.1732	0.029997
11	16	10.8361	173.3776	256	117.4211	4	2.161392	8.645569	16	10.8361	11.57109	-0.73499	0.540217
12	17	13.615	231.455	289	185.3682	5	4.940292	24.70146	25	13.615	12.29519	1.319809	1.741895
13	18	13.531	243.558	324	183.088	6	4.856292	29.13775	36	13.531	13.01929	0.511712	0.261849
合计	**156**	**112.7712**	**1485.04**	**2054**	**1083.376**	**0**	**0**	**131.7856**	182	**105.1183**	**112.7712**	≅ 0	**9.69281**

Example 3.6 ： 受教育程度(X)与时均工资(Y)
——基于表2-6的样本数据：参数估计值和回归方程

（思考）

- 公式1:（Favorite Five，FF形式）

$$\begin{cases}\hat{\beta}_2=\dfrac{n\sum X_iY_i-\sum X_i\sum Y_i}{n\sum X_i^2-\left(\sum X_i\right)^2}=\dfrac{13*1485.04-156*112.7712}{13*2054-(156)^2}=0.7240967\\ \hat{\beta}_1=\dfrac{n\sum X_i^2Y_i-\sum X_i\sum Y_iX_i}{n\sum X_i^2-\left(\sum X_i\right)^2}=\bar{Y}_i-\hat{\beta}_2\bar{X}_i=8.674708-0.72440967*12=-0.01445275\end{cases}$$

- 公式2：（离差形式）

$$\begin{cases}\hat{\beta}_2=\dfrac{\sum x_iy_i}{\sum x_i^2}=\dfrac{131.7856}{182}=0.7240967\\ \hat{\beta}_1=\bar{Y}_i-\hat{\beta}_2\bar{X}_i=8.674708-0.72440967*12=-0.01445275\end{cases}$$

（思考）

- 回归方程：

$$\hat{Y}_i=\hat{\beta}_1+\hat{\beta}_2X_i=-0.01445275+0.7240967X_i$$

Example 3.6 ： 受教育程度(X)与时均工资(Y)
——基于表2-6的样本数据：参数估计值和样本回归线

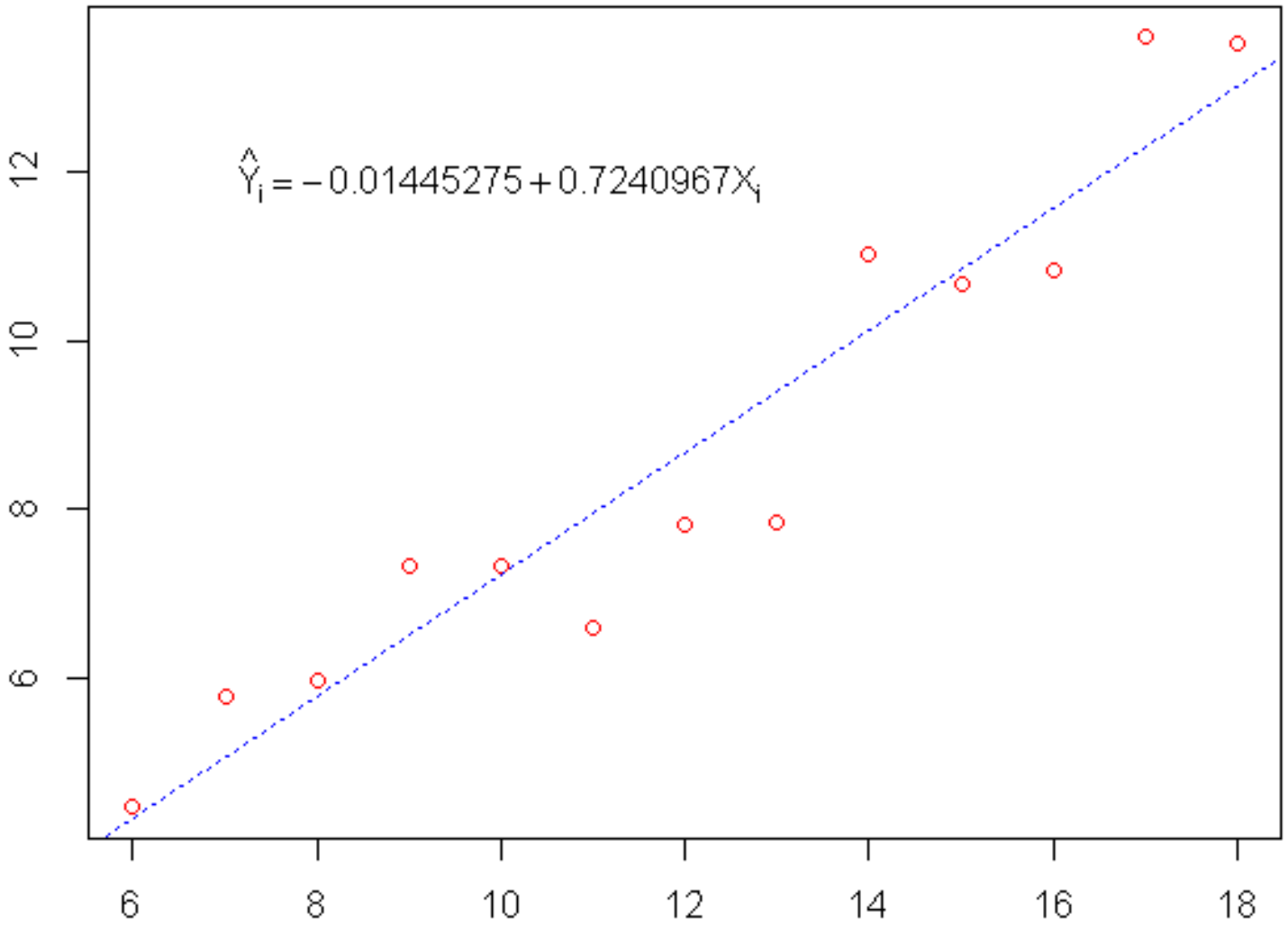


Figure 3-11

（思考）

Example 3.6： 受教育程度(X)与时均工资(Y)
——基于表2-6的样本数据:$\hat{\sigma}^2$、$var(\beta_i)$、$se(\beta_i)$、

回归误差方差

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{(n-2)} = \frac{9.83017}{11} = 0.893625$$

回归误差标准差se

$$\hat{\sigma} = \sqrt{\frac{\sum e_i^2}{(n-2)}} = 0.945332$$

（式 3.5.8）
残差-离差形式

$$\text{var}(\hat{\beta}_2) = \frac{\hat{\sigma}^2}{\sum x_i^2} = \frac{0.893652}{182} = 0.00491; \quad se(\hat{\beta}_2) = 0.070072$$

（式 3.5.6/7）
系数-离差形式

$$\text{var}(\hat{\beta}_1) = \frac{\sum X_i^2}{n\sum x_i^2}\hat{\sigma}^2 = \frac{2054*0.893652}{13*182} = 0.775808; \quad se(\hat{\beta}_1) = 0.8808$$

（式 3.5.8）
离差形式

Example 3.6 ： 受教育程度(X)与时均工资(Y)
——基于表2-6的样本数据：离差分解和可决系数

$$RSS=\sum e_i^2=9.69281$$

$$TSS=\sum (Y_i-\bar{Y})^2=105.1183$$

公式1:（Favorite Five，FF形式）

（式 3.5.8）
残差-离差形式

$$r^2=\frac{ESS}{TSS}=1-\frac{RSS}{TSS}=1-\frac{\sum e_i^2}{\sum (Y_i-\bar{Y})^2}=1-\frac{9.69281}{105.1183}=0.9077914$$

（式 3.5.6/7）
系数-离差形式

$$r^2=\hat{\beta}_2^2\left(\frac{\sum x_i^2}{\sum y_i^2}\right)=0.7240967^2\left(\frac{182}{105.1183}\right)=0.9077914$$

（式 3.5.8）
离差形式

$$r^2=\frac{\left(\sum x_i y_i\right)^2}{\sum x_i^2\cdot\sum y_i^2}=\frac{(131.7856)^2}{182*105.1183}=0.9077914$$

Excise 3.27：课堂作业
——基于蒙特卡罗法的重复抽样数据

本节内容：

§2.4 经典正态线性回归模型

- §2.4.1 干扰项 μ_i 的概率分布
- §2.4.2 干扰项 μ_i 的正态性假设
- §2.4.3 正态性假设：OLS估计量的性质
- §2.4.4 极大似然估计法(ML)：简述

经典正态线性回归模型（CNLRM）
——干扰项 μ_i的正态性假设：引子

(L-SRF-s)

(L-PRF-s)

- 估计SRF→推断PRF：

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

CLRM假设下：OLS方法得到的$\hat{\beta}_2$和$\hat{\beta}_1$已经是BLUE了

CLRM假设下：μ_i的期望值为零，不相关的，且有一个不变方差$\sigma^2(\mu_i)$。

$$Y_i = E(Y \mid X_i) + u_i = \beta_1 + \beta_2 X_i + \mu_i$$

- 问题：
  - 一个样本怎样才推断总体PRF？
  - $\hat{\beta}_2$、$\hat{\beta}_1$和$\hat{\sigma}^2(e_i)$因为样本变化而变化(随机变量)！
- μ_i对于估计SRF→推断PRF的重要性：

$$\hat{\beta}_2 = \sum\left(\frac{x_i}{\sum x_i^2}\cdot Y_i\right) = \sum k_i \cdot Y_i = \sum k_i \cdot (\beta_1 + \beta_2 X_i + \mu_i)$$

$$\hat{\beta}_1 = \sum\left[\left(\frac{1}{n} - \bar{X}k_i\right)\cdot Y_i\right] = \sum w_i \cdot Y_i = \sum w_i \cdot (\beta_1 + \beta_2 X_i + \mu_i)$$

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

经典正态线性回归模型（CNLRM）
——干扰项 μ_i的正态性假设：表达

- 经典正态线性回归模型(classical normal linear regression model , CNLRM)
  - 在经典线性回归模型(CLRM)假设中再增加**干扰项 μ_i的正态性假设：**

（式 4.2.1）均值：$E(\mu_i)=0$

（式 4.2.2）方差：$E(\mu_i^2)=\sigma^2$

（式 4.2.3）协方差：$\text{cov}(\mu_i,\mu_j)$: $E\left\{\left[\mu_i-\text{E}(\mu_i)\right]\cdot\left[\mu_j-\text{E}(\mu_i)\right]\right\}=E(\mu_i\cdot\mu_j)=0,\ i\neq j$

（式 4.2.4）正态性：$\mu_i\sim N(0,\sigma^2);\quad \mu_j\sim N(0,\sigma^2);$

（式 4.2.5）独立正态性：$\mu_{(i,j)}\sim \text{NID}(0,\sigma^2)$

– 其中NID代表*正态独立同分布*!

经典正态线性回归模型（CNLRM）
——OLS估计量的性质：优良特性

- CNLRM下，OLS估计量有如下统计性质：
  - ① 无偏性
  - ② 有效性（方差最小）
  - ③ 一致性（收敛到它们的总体参数上）
  - ④$\hat{\beta}_2$(它是μ_i的线性函数)是正态分布的
  - ⑤$\hat{\beta}_1$(它是μ_i的线性函数)是正态分布的
  - ⑥$(n-2)\hat{\sigma}^2/\sigma^2$服从自由度为(n-2)的$\chi^2$分布
  - ⑦$(\hat{\beta}_2, \hat{\beta}_1)$的分布独立$\hat{\sigma}^2$
  - ⑧$\hat{\beta}_2$和$\hat{\beta}_1$在所有无偏估计中，无论是线性还是非线性，都有最小的方差。也即，它们是**最有无偏估计量**（best unbiased estimators, BUE）。

经典正态线性回归模型（CNLRM）
——OLS估计量的性质：$\hat{\beta}_2$和$\hat{\beta}_1$是正态分布(1/2)

- ④ $\hat{\beta}_2$（它是μ_i的线性函数）是正态分布的

$$E(\hat{\beta}_2)=\beta_2,\quad var(\hat{\beta}_2)=\sigma^2_{\hat{\beta}_2}=\frac{\sigma^2}{\sum x_i^2}$$

$$\hat{\beta}_2 \sim N(\beta_2,\sigma^2_{\hat{\beta}_2})$$

σ^2

$$Z=\frac{\hat{\beta}_2-\beta_2}{\sigma_{\hat{\beta}_2}} \sim N(0,1)$$

- ⑤ $\hat{\beta}_1$（它是μ_i的线性函数）是正态分布的

$$E(\hat{\beta}_1)=\beta_1,\quad var(\hat{\beta}_1)=\sigma^2_{\hat{\beta}_1}=\frac{\sum X_i^2}{n\sum x_i^2}\cdot\sigma^2$$

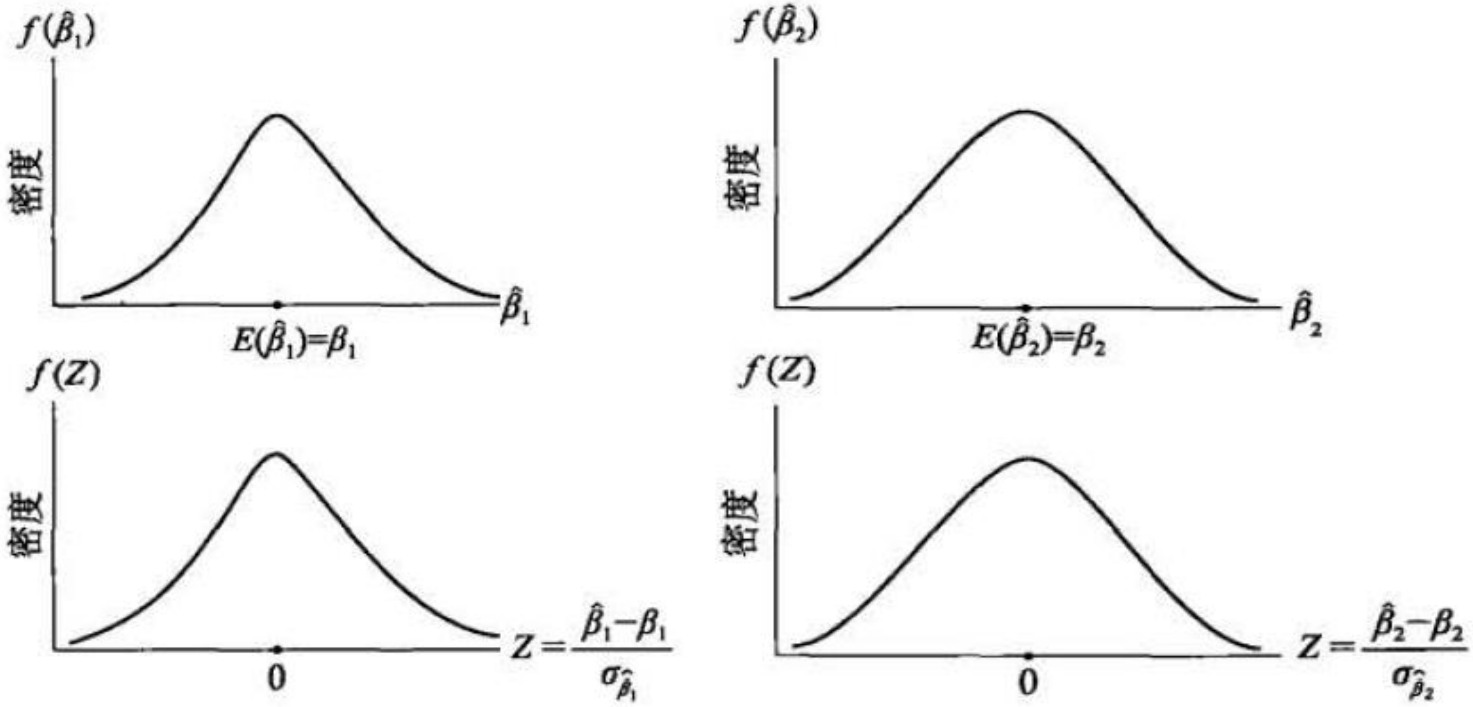
$$\hat{\beta}_1 \sim N(\beta_1,\sigma^2_{\hat{\beta}_1})$$

σ^2

$$Z=\frac{\hat{\beta}_1-\beta_1}{\sigma_{\hat{\beta}_1}} \sim N(0,1)$$

经典正态线性回归模型（CNLRM）
——OLS估计量的性质：$\hat{\beta}_2$和$\hat{\beta}_1$是正态分布(2/2)

图4-1 $\hat{\beta}_2$和$\hat{\beta}_1$的概率分布



- 另外：Y_i（它是μ_i的线性函数）是正态分布的

$$E(Y_i \mid X_i) = \beta_1 + \beta_2 X_i, \quad \text{var}(Y_i) = \sigma^2$$

$$Y_i \sim N(\beta_1 + \beta_2 X, \ \sigma^2)$$

极大似然估计法(ML)
——ML的思想

- 极大似然估计法(ML)的原理：
  - 极大似然估计法(maximum likelihood, ML)：
    - 是由Fisher提出的一种参数估计方法
    - 基本思想：设总体分布的函数形式已知，但有未知参数θ，θ可以取很多值，在θ的一切可能取值中选一个使样本观察值出现的概率为最大的θ值作为θ的估计值，记$\hat{\theta}$，并称之为的极大似然估计值。这种求估计量的方法称为极大似然估计法。
  - 似然函数表达式：
    - 设总体X的概率密度函数为$f(X;\theta)$，其中θ为待估计参数。对于从总体中取得的样本观察值$X_1, X_1, \ldots, X_n$，其联合密度函数为$\prod f(X_i;\theta)$，它是参数θ的函数，称之为θ的似然函数，记为L(θ)：

$$L(\theta)=\prod_{i=1}^{n} f(x_i;\theta)$$

极大似然估计法(ML)
——ML估计法与OLS估计法的关系：在本书中的意义

- 极大似然估计法(ML)比较复杂，我们仅需知道：
  - [CNLRM假设中]在干扰项 μ_i 的正态性的假设下：
    - 回归系数β的ML估计量和OLS估计量是相同的——无论是一元回归还是多元回归!
    - 对于 σ^2 的估计，其ML估计量为 $\sum e_i^2/n$，是有偏的；而其OLS估计量为 $\sum e_i^2/(n-2)$ 是无偏的。
    - 关于 σ^2 的两种估计量，随着样本容量n的增大，两者将趋于相等！
- 启示：OLS方法真好！

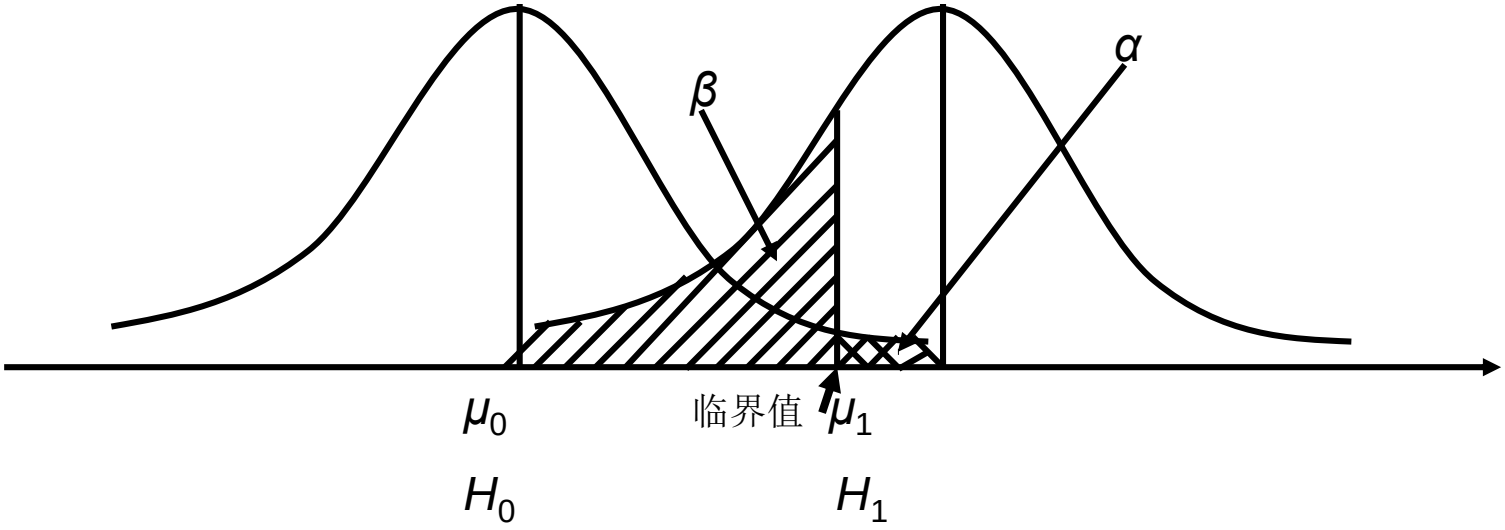
本节内容：

§2.5 一元线性回归模型：区间估计和假设检验

- §2.5.1 统计学的预备知识
- §2.5.2 区间估计：一些基本思想
- §2.5.3 回归系数β_i的置信区间
- §2.5.4σ^2的置信区间(*不要求掌握)
- §2.5.5 假设检验:概述
- §2.5.6 假设检验：置信区间的方法
- §2.5.7 假设检验：显著性检验法
- §2.5.8 假设检验：实际操作
- §2.5.9 回归分析与方差分析
- §2.5.10 回归分析应用：预测问题
- §2.5.11 报告回归分析结果

统计学的预备知识*
——自己去复习

- 显著性水平 α
- 统计检验的功效（能力）
- 置信度
- 置信区间
- 第I类错误：弃真错误$\alpha = P(Z > Z_0 | H_0\text{为真})$
- 第II类错误：取伪错误$\beta = P(Z \le Z_0 | H_1\text{为真})$



区间估计：*
——基本概念（自己去巩固）

（式 5.2.1）

- 点估计和区间估计

$$\Pr(\hat{\beta}_2 - \delta \le \beta_2 \le \hat{\beta}_2 + \delta) = 1 - \alpha$$

- 随机区间(random interval）：　$(\hat{\beta}_2 - \delta, \hat{\beta}_2 + \delta)$
- 置信区间(confidence interval)：$\hat{\beta}_2 - \delta \le \beta_2 \le \hat{\beta}_2 + \delta$
- 显著性水平(level of significance)：　α
- 置信度或置信系数(confidence coefficient):　$1-\alpha$
- 置信限（confidence limits）或临界值（critical values）
- 置信上限（lower confidence limit）
- 置信下限（uper confidence limit）

区间估计：*
——注意的几个问题（自己去巩固）

（式 5.2.1）

$$\Pr(\hat{\beta}_2 - \delta \le \beta_2 \le \hat{\beta}_2 + \delta) = 1 - \alpha$$

- 陈述问题：
  - β 2落入给定界限内的概率是1-α。 （X）？？
  - 使用我们的方法构造出来的区间包含β_2的概率为1-α。
  - 抽样层面来理解：从重复多次抽样中来看，平均起来这些区间将有(1-α)*100%的可能包含着**参数**的真值。
- 我们构造的区间是只是随机区间！（？）
  - 对于计算出的参数估计值而言，得到的区间中要么包含参数真值要么不包含。概率为0或1！
  - 例如：对于95%置信区间的$0.4268 \le \beta_2 \le 0.5914$而言，不能说这个区间包含真实的$\beta_2$的概率是95%。这个概率不是1就是0。

区间估计： *
——注意的几个问题（自己去巩固）

（式 5.2.1）

- 两个游戏：
  - 掷硬币
  - 套圈

  请问：区间估计更象哪一个?
- 置信区间的两个特点：
  - 位置的随机性
  - 长度的随机性

回归系数β_i的置信区间
——以β_2为例：构建t分布

（式 5.3.1）

$$Z=\frac{(\hat{\beta}_2-\beta_2)}{\sqrt{\mathrm{var}(\hat{\beta}_2)}}=\frac{(\hat{\beta}_2-\beta_2)}{\sqrt{\sigma^2_{\hat{\beta}_2}}}=\frac{\hat{\beta}_2-\beta_2}{\sigma_{\hat{\beta}_2}}=\frac{(\hat{\beta}_2-\beta_2)}{\sqrt{\frac{\sigma^2}{\sum x_i^2}}}$$

Z □ N(0,1)

（式 5.3.2）

$$T=\frac{(\hat{\beta}_2-\beta_2)}{\sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2}}}=\frac{\hat{\beta}_2-\beta_2}{\sqrt{S^2_{\hat{\beta}_2}}}=\frac{\hat{\beta}_2-\beta_2}{S_{\hat{\beta}_2}}$$

T □ $t(n-2)$

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

（式 5.3.3）

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

回归系数β_i的置信区间
——以β_2为例：构建t分布

（式 5.3.4）

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

（式 5.3.5）

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

（式 5.3.6）

β_2的$100(1-\alpha)\%$置信区间为：$\hat{\beta}_2 \pm t_{\alpha/2} \cdot S_{\hat{\beta}_2}$

（式 5.3.8）

类似地推理有β_1的$100(1-\alpha)\%$置信区间为：$\hat{\beta}_1 \pm t_{\alpha/2} \cdot S_{\hat{\beta}_1}$

回归系数β_i的置信区间
——示例：教育程度与时均工资回归

- 给定：$\alpha = 0.05, (1-\alpha)100\% = 95\%$
- 查t表：$t_{\alpha/2}(n-2) = t_{0.05/2}(11) = 2.201$
- 我们之前已算出：

$$\hat{\beta}_1 = -0.014; S_{\hat{\beta}_1} = 0.8808; \hat{\beta}_2 = 0.724; S_{\hat{\beta}_2} = 0.070072$$

那么，β_2的95%置信区间分别为：

$$\hat{\beta}_2 - t_{\alpha/2} \cdot S_{\hat{\beta}_2} \le \beta_2 \le \hat{\beta}_2 + t_{\alpha/2} \cdot S_{\hat{\beta}_2}$$

$$0.724 - 2.201*0.07 \le \beta_2 \le 0.724 + 2.201*0.07$$

$$0.57 \le \beta_2 \le 0.878$$

同理：β_1的95%置信区间为：

$$-1.8871 \le \beta_1 \le 1.8583$$

σ^2的置信区间
——构建χ^2分布*(不要求掌握)

（式 5.4.1）

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

$$\chi^2 \sim \chi^2(n-2)$$

（式 5.4.2）

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

（式 5.4.3）

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

假设检验
——概览：基本原理和思路*

（定义）

- 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）。

（定义）

- 虚拟假设(null hypothesis) ——H_0
  - 指定或声称的假设
  - 它是一个等待被挑战的“靶子”！ “稻草人”！

（定义）

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

- 统计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）

（思考）

讨论1：参数的区间估计和假设检验有什么区别和联系？

假设检验：置信区间的方法（1/2）
——双侧检验*

（定义）

- 双侧或双尾检验（Two-sided or Two-Tail Test）：

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

简单假设

复合假设

（约定）

- 假设检验目的：估计的$\hat{\beta}_2$是否与上述H_0相容？
- 置信区间的检验方法——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。

统计不显著！

统计显著！

（Figure 5-2）

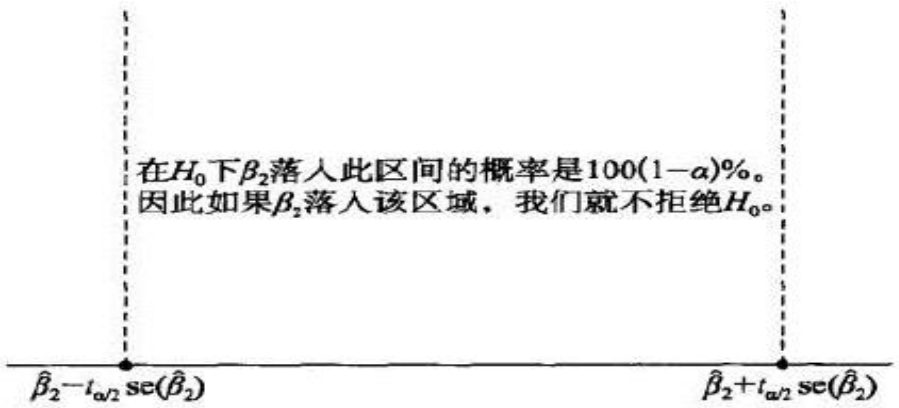


图 5—2　β_2 的一个 $100(1-\alpha)\%$置信区间

（思考）

讨论1：程度判断问题：“显著的”、“中度显著的”、“高度显著的”？

假设检验：置信区间的方法（2/2）
——单侧检验*

（定义）

- 单侧或单尾检验(One-Sided or One-Tail Test):
  - 当我们有着很强的理论支撑或者先验性预期时，可以把备择假设H1取为单侧的或单向的。

$$H_0 : \beta_2 \le 0.5$$ 简单假设

$$H_1 : \beta_2 > 0.5$$ 简单假设

（思考）

  - 对于这种单尾检验，最好的方法是用下面要讲到的显著性检验方法。（？）

假设检验：置信区间的方法*

——示例：3.6教育程度与时均工资

β_2假设：

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

β_1假设：

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

- 给定：$\alpha = 0.05, (1-\alpha)100\% = 95\%$
- 查t表：$t_{\alpha/2}(n-2) = t_{0.05/2}(11) = 2.201$
- 我们之前已算出参数估计值及样本标准差：

$$\hat{\beta}_2 = 0.724; \ S_{\hat{\beta}_2} = 0.070072; \ \hat{\beta}_1 = -0.014; \ S_{\hat{\beta}_1} = 0.8808$$

β_1和β_2的95%置信区间分别为：

$$0.57 \leq \beta_2 \leq 0.878; \ -1.8871 \leq \beta_1 \leq 1.8583$$

（思考）

➢ 那么我们可以做出如下检验判断：

对β_2假设：拒绝原假设H_0，接受H_1。认为，长期来看很多个区间(0.57，0.878)有95%的可能性不包含0.5($\beta_2 \neq 0.5$)

对β_1假设：不能拒绝原假设H_0。认为，长期来看很多个区间(-1.8871，1.8583)有95%的可能性不能拒绝$\beta_1 = 0$)

假设检验：显著性检验法
——原理和思路

（定义）

- 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

（思考）

- 关键思路：
  - 找到一个适合的检验**统计量**(test statistic)
  - 知道该**统计量**在H_0下的抽样分布(pdf)
  - 计算**样本统计量的值**
  - 查表找出给定显著水平α下的**理论统计量**的**临界值**
  - 比较**样本统计量值**和该**临界值**的大小
  - 做出拒绝还是接受H_0的判断！

用样本数据可快速估计出来

t、χ^2、F、正态分布等

假设检验：显著性检验法（2/2）很重要！

——检验回归系数$\hat{\beta}_i$的显著性：t检验

0.写出模型

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

1.提出假设

对β_2：$H_0 : \beta_2 = 0$，$H_1 : \beta_2 \neq 0$

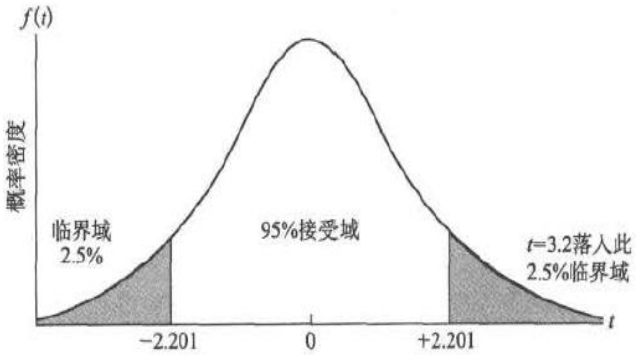
对β_1：$H_0 : \beta_2 = 0$，$H_1 : \beta_2 \neq 0$

2.根据H_0和样本数据，计算样本统计量T*

$$T = \frac{\hat{\beta}_2 - \beta_2}{S_{\hat{\beta}_2}} = \frac{(\hat{\beta}_2 - \beta_2)}{\hat{\sigma} \Big/ \sqrt{\sum x_i^2}} \Rightarrow T^* = \frac{\hat{\beta}_2}{S_{\hat{\beta}_2}} = \frac{\hat{\beta}_2}{\hat{\sigma} \Big/ \sqrt{\sum x_i^2}}$$

$$T \sim t(n-2)$$

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



3.比较T*和查表的$t_{\alpha/2,n-2}$

4.得出t检验结论

a.若$|T^*| > t_{\alpha/2,n-2}$，则β_2的t检验结果显著，表明应拒绝H_0,接受H_1,认为β_2显著不等于0

b.若$|T^*| < t_{\alpha/2,n-2}$，则β_2的t检验结果不显著，表明不能拒绝H_0,认为不能拒绝β_2等于0。

假设检验：实际操作
——“拒绝”或“接受”假设的含义

β_2假设：

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

- “拒绝”或“接受”假设的含义：

$$T^* = \frac{0.24}{0.007} = 0.3438$$

$$t_{\alpha/2, n-2} = 2.201$$

（思考）

- 若t检验不显著，判断为：“接受H0”
  - 只表明不能拒绝H_0，并不表明H_0毫无疑问是真！
  - 正如，法庭宣告某一判决为“无罪”而不为“清白”！

假设检验：实际操作

——"零"虚拟假设与"2-t"经验法则

β_2假设：

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

- "零"虚拟假设与"2-t" 经验法则：

当自由度$f \geq 20$, 显著水平$\alpha = 0.05$，

如果根据H_0算出：

$$|T^*| = \left|\frac{\hat{\beta}_2 - \beta_2}{S_{\hat{\beta}_2}}\right| = \left|\frac{\hat{\beta}_2}{S_{\hat{\beta}_2}}\right| > 2$$

$t_{0.05/2}(20) = -2.085963$

（思考）

（双侧检验下）则可以认为拒绝H0，接受H1！

假设检验：实际操作
——构造虚拟假设和对立假设

- 构造虚拟假设和对立假设
  - 如何有效构建H0和H1？？？？

β_2假设：

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

（思考）

- 证券组合理论：风险越大，收益越高，系数应该为正

$$Ei = \beta_1 + \beta_2 \sigma_i$$

- 货币需求弹性：典型地处于(0.7, 1.3)之间！

假设检验：实际操作

——显著性水平：选择α，或者确定p值*

（思考）

- 选择显著性水平α
  - 犯错误类型：
    - 第I类错误：弃真错误$\alpha = P(Z > Z_0 | H_0$为真$)$
    - 第II类错误：取伪错误$\beta = P(Z \leq Z_0 | H_1$为真$)$
    - [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α通常都固定在1%、5%或者10%的水平上?
    - 约定而已，并非神圣不可改变！
    - 如何改变？？
- 精确的显著性水平：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
——统计意义、经济意义及检验方法选择

- 统计显著性与实际显著性
  - 例子说明：边际消费倾向(MPC)
    - 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$\widehat{MPC}$的95%置信区间为(0.7129, 0.7306)，而当样本表明$\widehat{MPC}$为0.74(即乘数为3.84)，你怎样抉择！！！

- 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：β_2=0 显然荒谬的！

（思考）

回归分析与方差分析
——概念与思想

- 方差分析（analysis of variance，ANOVA）
  - 定义：对TSS的构成部分进行研究
  - 目的：分析和检验某种或多种因素（或自变量）的变化对试验结果（或因变量）的观测数据是否有显著的影响。

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2 + 2\sum \hat{y}_i e_i = \sum \hat{y}_i^2 + \sum e_i^2 = \hat{\beta}_2^2 \sum x_i^2 + \sum e_i^2$$

TSS = ESS + RSS

（思考）

总离差平方和 = 回归平方和 + 剩余平方和

  - 概率分布：在数理统计中，可以证明：

$$TSS \sim \chi^2(n-1);\ ESS \sim \chi^2(1);\ RSS \sim \chi^2(n-2)$$

回归分析与方差分析
——双变量方差分析表

- 双变量方差分析表（ANOVA）(做一遍)

变异来源	平方和SS	自由度df	均方和MSS
回归部分	**ESS** $\sum \hat{y}_i^2 = \hat{\beta}_2^2 \sum x_i^2$	1	MSS_ESS $\sum \hat{y}_i^2 = \hat{\beta}_2^2 \sum x_i^2$
剩余部分	**RSS** $\sum e_i^2$	n-2	**MSS_RSS** $\frac{\sum e_i^2}{n-2} = \hat{\sigma}^2$
总和	**TSS** $\sum y_i^2$	n-1	**MSS_TSS** $\frac{\sum y_i^2}{n-1}$

（思考）

回归分析与方差分析
——方差比F和F检验

- 构造F统计量：$\boldsymbol{F \sim F(1, n-2)}$

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

  - 下面两式可证：

（式5.9.2）

$$E(\hat{\beta}_2^2 \sum x_i^2) = \sigma^2 + \beta_2^2 \sum x_i^2$$

证明见下页：

（式5.9.3）

$$E\left[\frac{\sum e_i^2}{n-2}\right] = E(\hat{\sigma}^2) = \sigma^2$$

我们已证明：

- F检验：$H_0: \beta_2 = 0$；$H_1: \beta_2 \neq 0$
  - 如果由样本数据计算的$\boldsymbol{F^*} > F_\alpha$，则拒绝$H_0$，表示在显著性水平之下，该一元回归模型是统计上显著的

附录2.5.9：证明（式5.9.2）：

$$
\begin{aligned}
E(\hat{\beta}_2^{\,2}\sum x_i^{\,2}) &= \sum x_i^{\,2}E(\hat{\beta}_2^{\,2}) \\
&= \sum x_i^{\,2}\mathrm{E}\left[\beta_2^{\,2} + (\sum k_i u_i)^2 + 2\beta_2 \sum k_i u_i\right] \\
&= \sum x_i^{\,2}\left[\beta_2^{\,2} + E(\sum k_i u_i)^2\right] \\
&= \sum x_i^{\,2}\left[\beta_2^{\,2} + \frac{\sigma^2}{\sum x_i^{\,2}}\right] \\
&= \sigma^2 + \beta_2^{\,2}\sum x_i^{\,2}
\end{aligned}
$$

$$
\begin{aligned}
\hat{\beta}_2 &= \sum k_i Y_i \\
&= \sum k_i(\beta_1 + \beta_2 X_i + u_i) \\
&= \beta_2 + \sum k_i u_i
\end{aligned}
$$

$$
\begin{aligned}
E(\sum k_i u_i)^2 &= E(k_1u_1 + k_2u_2 + \cdots + k_nu_n)^2 \\
&= k_1^{\,2}Eu_1^{\,2} + k_2^{\,2}Eu_2^{\,2} + \cdots + k_n^{\,2}Eu_n^{\,2} \\
&= \sigma^2\sum k_i^{\,2} = \frac{\sigma^2}{\sum x_i^{\,2}}
\end{aligned}
$$

回归分析与方差分析

—模型整体显著性的F检验（很重要！）

0.写出模型

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

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + e_i$$

1.提出斜率系数的联合假设

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

$$H_0 : \beta_2 = \beta_3 = 0;$$
$$H_1 : \beta_2 \text{和} \beta_3 \text{不全为0}$$

2.计算样本统计量F*

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

3.比较F*和查表的$F_{\alpha,(f_1,f_2)}$

4.得出F检验结论

a.若$F^* > F_{\alpha,(f_1,f_2)}$，则回归模型的整体验结果显著，表明应拒绝$H_0$,接受$H_1$,可以认为回归模型是有意义的。

b.若$F^* < F_{\alpha,(f_1,f_2)}$，则回归模型的整体验结果不显著，表明不能拒绝$H_0$,从而不能认为回归模型是有意义的。

回归分析与方差分析
——F检验和t检验的异同及联系

- 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$H_0: \beta_2 = 0$；$F_{1,n-2} = t^2_{n-2}$（式5.9.2）
- F检验与t检验的不同：
  - 检验目的不同
    - F检验：模型的显著性；
    - t检验：系数的显著性
  - 假设的提出不同：（式5.9.3）
    - F检验：斜率系数联合假设$H_0: \beta_2 = \beta_3 = \cdots = \beta_k = 0$
    - t检验：回归系数分别假设$H_0: \beta_1 = 0$或$H_0: \beta_2 = 0$
  - 检验原理的不同：
    - F检验：构造F统计量；
    - t检验：构造T统计量

回归分析与方差分析
——示例：教育程度与时均工资

Table 2-6

表2-6 不同受教育程度的平均小时工资

分组	读书年数	时均工资$	人数
1	6	4.4567	3
2	7	5.77	5
3	8	5.9787	15
4	9	7.3317	12
5	10	7.3182	17
6	11	6.5844	27
7	12	7.8182	218
8	13	7.8351	37
9	14	11.0223	56
10	15	10.6738	13
11	16	10.8361	70
12	17	13.615	24
13	18	13.531	31
			528

此数据源自：1985年5月的美国人口普查

Figure 2-6

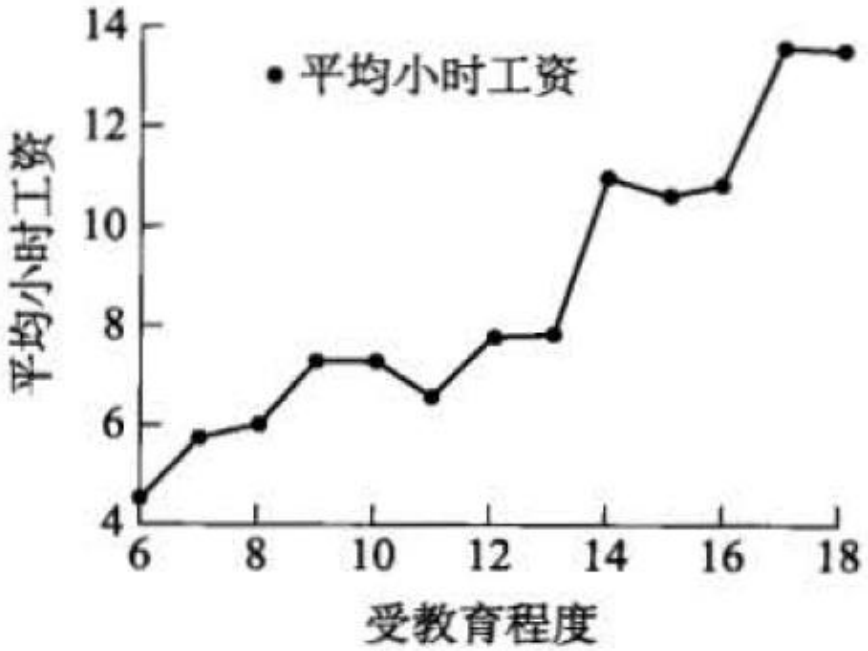


图2-6 平均小时工资与受教育程度之间的关系

思考：

回归分析与方差分析
——示例：教育程度与时均工资

1	样本	教育年数	时平均工资										
2	n	X	Y	XY	X2	Y2	x	y	xy	x2	y2	Y_reg	ei	ei2
3	1	6	4.4567	26.7402	36	19.8622	-6	-4.2180	25.3080	36	17.7916	4.3301	0.1266	0.0160
4	2	7	5.7700	40.3900	49	33.2929	-5	-2.9047	14.5235	25	8.4373	5.0542	0.7158	0.5123
5	3	8	5.9787	47.8296	64	35.7449	-4	-2.6960	10.7840	16	7.2685	5.7783	0.2004	0.0402
6	4	9	7.3317	65.9853	81	53.7538	-3	-1.3430	4.0290	9	1.8037	6.5024	0.8293	0.6877
7	5	10	7.3182	73.1820	100	53.5561	-2	-1.3565	2.7130	4	1.8401	7.2265	0.0917	0.0084
8	6	11	6.5844	72.4284	121	43.3543	-1	-2.0903	2.0903	1	4.3694	7.9506	-1.3662	1.8665
9	7	12	7.8182	93.8184	144	61.1243	0	-0.8565	0.0000	0	0.7336	8.6747	-0.8565	0.7336
10	8	13	7.8351	101.8563	169	61.3888					.7049	9.3988	-1.5637	2.4452
11	9	14	11.0223	154.3122	196	121.491					.5112	10.1229	0.8994	0.8089
12	10	15	10.6738	160.1070	225	113.930					.9964	10.8470	-0.1732	0.0300
13	11	16	10.8361	173.3776	256	117.421					.6716	11.5711	-0.7350	0.5402
14	12	17	13.6150	231.4550	289	185.368					.4065	12.2952	1.3198	1.7419
15	13	18	13.5310	243.5580	324	183.088					.5836	13.0193	0.5117	0.2618
16	合计	156	112.7712	836.6494	2054	1083.37					.1183	112.7712	0	9.6928
17	均值	12	8.6747											

18	$t_{0.05/2}(11)$	2.201
19	β1	-0.0145
20	β2	0.7241
21	TSS	105.1183
22	RSS	9.6928
23	ESS	95.4255
24	r2	0.9078
25	sigma_2	0.8812
26	Se_β2	0.0696
27	Se_β1	0.8746
28	t_β2	10.4065
29	t_β1	-0.0165

回归分析与方差分析
——示例：教育程度与时均工资

表 5—4　　工资—受教育程度一例的 ANOVA 表

变异来源	SS	df	MSS	
回归部分（ESS）	95.425 5	1	95.425 5	$F=\frac{95.425\ 5}{0.881\ 1}$
剩余部分（RSS）	9.692 8	11	0.881 1	=108.302 6
TSS	105.118 3	12		

（思考）

- $H_0: \beta_2 = 0$；$H_1: \beta_2 \neq 0$
  - F检验：$F^* = 108.3026$

    $F_\alpha(1, n-2) = F_{0.05}(1,11)=4.844336$

    $F_\alpha(1, n-2) = F_{0.01}(1,11)=9.646034$

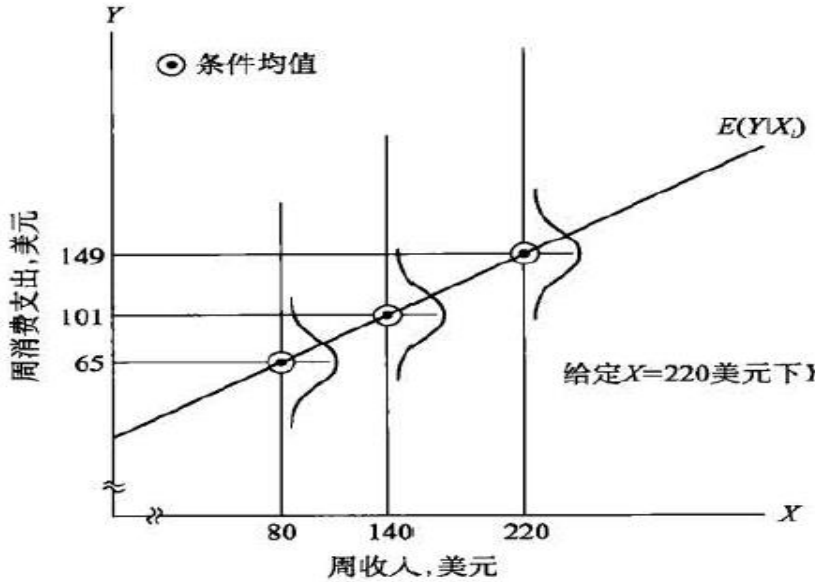
    由于$F^* > F_\alpha$，拒绝H_0，接受H_1　$(t^*)^2 = (10.41)^2 = 108.3681 \approx F^*$
  - t检验：

    $$t^* = \frac{\hat{\beta}_2}{se(\hat{\beta}_2)} \frac{0.724}{0.07} = 10.41 \qquad \left|t_{\alpha/2}(n-2)\right| = \left|t_{0.05/2}(11)\right| = 2.201$$

    由于$t^* > |t_{\alpha/2}|$，拒绝H_0，接受H_1

回归分析的应用
——预测问题：概览

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i = -0.0144 + 0.724 X_i$$



$$\hat{Y}_i \to \mathrm{E}(Y_i \mid X_i)$$

$$\hat{Y}_i \to Y_i \mid X_i$$

（思考）

- 均值预测(mean prediction)：
  - 给定X_0，预测Y的条件均值$\mathrm{E}(Y|X = X_0)$
- 个值预测(individual prediction)：
  - 给定X_0，预测对应于X_0 的Y的个别值 $(Y_0|X = X_0)$

均值预测(mean prediction)
——原理和思想(1/2)：$\hat{Y}_0$是正态分布

（思考）

- 给定X_0=20，Y的条件均值$E(Y|X_0 = 20)$的点估计为：

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0 = -0.0144 + 0.724 \times 20 = 14.4656$$

  - 可以证明：点估计$\hat{Y}_0$是真值$E(Y_0|X_0 = 20)$的一个BLUE

- 而且$\hat{Y}_0$具有如下正态分布（证明略）：

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

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

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

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

均值预测(mean prediction)
——原理和思想(2/2):t分布和$\hat{Y}_0$的置信区间

- 对$\hat{Y}_0$构造t统计量：

$$t=\frac{\hat{Y}_0-\mathrm{E}(\mathrm{Y}\mid \mathrm{X}_0)}{S_{\hat{Y}_0}}$$

$$S_{\hat{Y}_0}=\sqrt{\hat{\sigma}^2[\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum x_i^2}]}$$

- 均值$\mathrm{E}(\mathrm{Y}|\mathrm{X}_0=20)$的$(1-\alpha)100\%$置信区间为：

$$\Pr[\hat{Y}_0-t_{\alpha/2}\cdot S_{\hat{Y}_0}\le E(\mathrm{Y}\mid \mathrm{X}_0)\le \hat{Y}_0+t_{\alpha/2}\cdot S_{\hat{Y}_0}]=1-\alpha$$

（思考）

$$\Pr[\hat{\beta}_1+\hat{\beta}_2X_0-t_{\alpha/2}\cdot S_{\hat{Y}_0}\le E(\mathrm{Y}\mid \mathrm{X}_0)\le \hat{\beta}_1+\hat{\beta}_2X_0+t_{\alpha/2}\cdot S_{\hat{Y}_0}]=1-\alpha$$

均值预测(mean prediction)

——示例:教育程度和时均工资案例（$X_0 = 20$）

（思考）

- 根据表3-2的结果，可算出：

$$S^2_{\hat{Y}_0} = \hat{\sigma}^2[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}] = 0.8936(\frac{1}{13} + \frac{(20-13)^2}{182}) = 0.3826$$

$$S_{\hat{Y}_0} = 0.6185$$

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0 = -0.0144 + 0.724 \times 20 = 14.4656$$

- 那么，均值$E(Y|X_0 = 20)$的95%置信区间为：

$$\hat{Y}_0 - t_{\alpha/2} \cdot S_{\hat{Y}_0} \le E(Y|X_0=20) \le \hat{Y}_0 + t_{\alpha/2} \cdot S_{\hat{Y}_0}$$

$$14.4656 - 2.201*0.6185 \le E(Y|X_0=20) \le 14.4656 + 2.201*0.6185$$

$$13.1043 \le E(Y|X_0=20) \le 15.8260$$

个值预测(individual prediction)
——原理和思想(1/2)：Y_0是正态分布

- 给定X_0，预测对应于X_0的Y的个别值$(Y_0|X = X_0)$

$$Y_0 = \beta_1 + \beta_2 X_0 + u_0$$

- *易知*，$(Y_0|X = X_0)$的点估计值为：

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0$$
$$= -0.0144 + 0.724 \times 20$$
$$= 14.4656$$

  - 可证明，此点估计$\hat{Y}_0$是Y_0的ＢＬＵＥ（证明略）

- 易知，Y_0具有如下正态分布（很直接）：

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

（思考）

个值预测(individual prediction)
——原理和思想(2/2)：（$Y_0-\hat{Y}_0$）服从正态分布

（思考）

- 构造随机变量（$Y_0-\hat{Y}_0$）

$$Y_0 \sim \mathrm{N}(\beta_1+\beta_2 X_0, \sigma^2)$$

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

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

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

个值预测(individual prediction)
——原理和思想(2/2):构造t统计量和Y_0的置信区间

（思考）

- 对（$Y_0 - \hat{Y}_0$）是构造t统计量：

$$t^* = \frac{Y_0 - \hat{Y}_0}{S_{(Y_0 - \hat{Y}_0)}}$$

$$S_{(Y_0 - \hat{Y}_0)} = \sqrt{\hat{\sigma}^2 [1 + \frac{1}{n} + \frac{(X_0 - \bar{X})2}{\sum x_i^2}]}$$

- 真实值Y_0的$(1-\alpha)100\%$置信区间为：

$$\Pr\left[\hat{\beta}_1 + \hat{\beta}_2 X_0 - t_{\alpha/2} \cdot S_{(Y_0 - \hat{Y})} \le Y_0 \le \hat{\beta}_1 + \hat{\beta}_2 X_0 + t_{\alpha/2} \cdot S_{(Y_0 - \hat{Y})}\right] = 1 - \alpha$$

个值预测(mean prediction)
——示例:教育程度和时均工资案例（$X_0 = 20$）

（思考）

- 根据表3-2的结果，可算出：

$$S^2_{(Y_0-\hat{Y}_0)} = \hat{\sigma}^2[1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum x_i^2}]$$

$$=0.8936(1+\frac{1}{13}+\frac{(20-13)^2}{182})=1.2357$$

$$S_{(Y_0-\hat{Y}_0)} = 1.111586$$

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0 = -0.0144 + 0.724 \times 20 = 14.4656$$

- 那么，真实值$(Y_0|X_0 = 20)$的95%置信区间为：

$$\hat{Y}_0 - t_{\alpha/2} \cdot S_{(Y_0-\hat{Y}_0)} \le (Y_0|X_0=20) \le \hat{Y}_0 + t_{\alpha/2} \cdot S_{(Y_0-\hat{Y}_0)}$$

$$14.4656-2.201*1.111586 \le (Y_0|X_0=20) \le 14.4656+2.201*1.111586$$

$$12.019 \le (Y_0|X_0=20) \le 16.9122$$

预测(mean prediction)
——示例比较:教育程度和时均工资案例（置信带）

（定义）

- 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

$$\Pr[\hat{\beta}_1+\hat{\beta}_2X_0-t_{\alpha/2}\cdot S_{\hat{Y}_0}\le E(Y_0\mid X_0)\le \hat{\beta}_1+\hat{\beta}_2X_0+t_{\alpha/2}\cdot S_{\hat{Y}_0}]=1-\alpha$$

（图5-6）

$$S_{\hat{Y}_0}=\sqrt{\hat{\sigma}^2[\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum x_i^2}]}$$

$$\Pr\left[\hat{\beta}_1+\hat{\beta}_2X_0-t_{\alpha/2}\cdot S_{(Y_0-\hat{Y})}\le Y_0\le \hat{\beta}_1+\hat{\beta}_2X_0+t_{\alpha/2}\cdot S_{(Y_0-\hat{Y})}\right]=1-\alpha$$

$$S_{(Y_0-\hat{Y}_0)}=\sqrt{\hat{\sigma}^2[1+\frac{1}{n}+\frac{(X_0-\bar{X})2}{\sum x_i^2}]}$$

预测(mean prediction)
——示例比较:教育程度和时均工资案例（置信带）

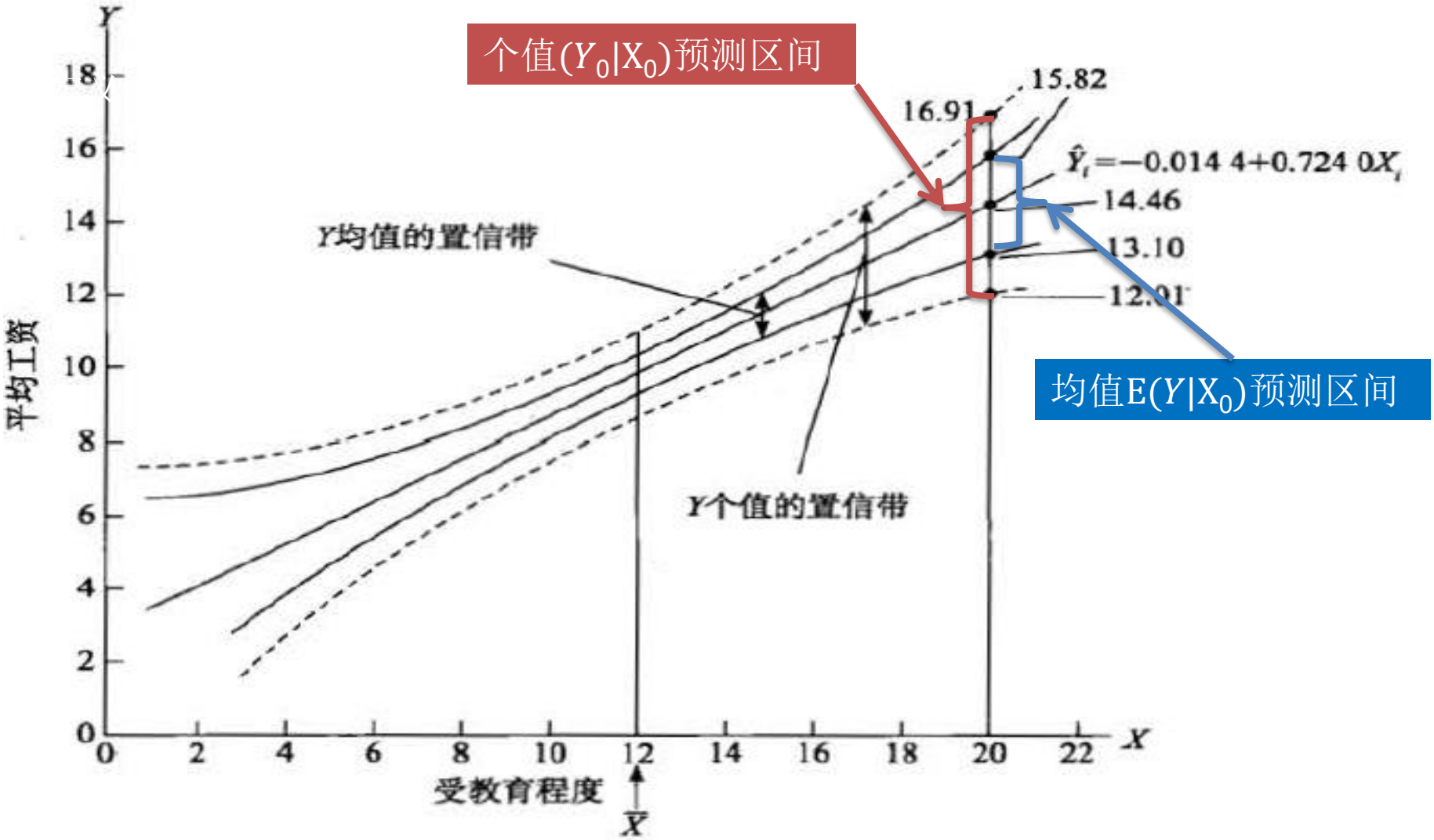


图 5—6　Y 均值与 Y 个值的置信带（域）

会看、会算
——案例1：受教育程度与时均工资

Table 2-6

表2-6 不同受教育程度的平均小时工资

分组	读书年数	时均工资$	人数
1	6	4.4567	3
2	7	5.77	5
3	8	5.9787	15
4	9	7.3317	12
5	10	7.3182	17
6	11	6.5844	27
7	12	7.8182	218
8	13	7.8351	37
9	14	11.0223	56
10	15	10.6738	13
11	16	10.8361	70
12	17	13.615	24
13	18	13.531	31
			528

此数据源自：1985年5月的美国人口普查

Figure 2-6

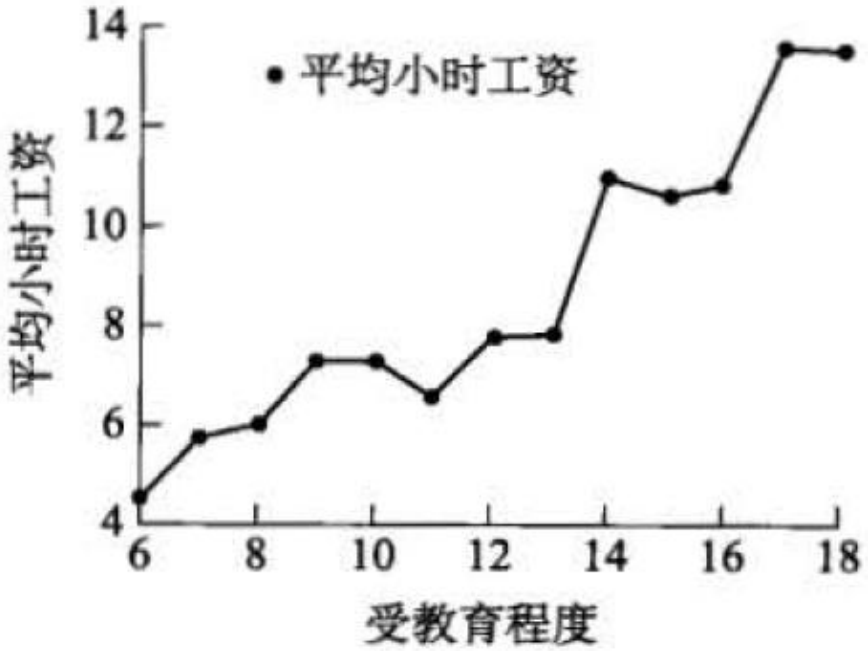


图2-6 平均小时工资与受教育程度之间的关系

思考：

会看、会算(手工计算)
——案例1：受教育程度与时均工资

1	样本	教育年数	时平均工资										
2	n	X	Y	XY	X2	Y2	x	y	xy	x2	y2	Y_reg	ei	ei2
3	1	6	4.4567	26.7402	36	19.8622	-6	-4.2180	25.3080	36	17.7916	4.3301	0.1266	0.0160
4	2	7	5.7700	40.3900	49	33.2929	-5	-2.9047	14.5235	25	8.4373	5.0542	0.7158	0.5123
5	3	8	5.9787	47.8296	64	35.7449	-4	-2.6960	10.7840	16	7.2685	5.7783	0.2004	0.0402
6	4	9	7.3317	65.9853	81	53.7538	-3	-1.3430	4.0290	9	1.8037	6.5024	0.8293	0.6877
7	5	10	7.3182	73.1820	100	53.5561	-2	-1.3565	2.7130	4	1.8401	7.2265	0.0917	0.0084
8	6	11	6.5844	72.4284	121	43.3543	-1	-2.0903	2.0903	1	4.3694	7.9506	-1.3662	1.8665
9	7	12	7.8182	93.8184	144	61.1243	0	-0.8565	0.0000	0	0.7336	8.6747	-0.8565	0.7336
10	8	13	7.8351	101.8563	169	61.3888					.7049	9.3988	-1.5637	2.4452
11	9	14	11.0223	154.3122	196	121.491					.5112	10.1229	0.8994	0.8089
12	10	15	10.6738	160.1070	225	113.930					.9964	10.8470	-0.1732	0.0300
13	11	16	10.8361	173.3776	256	117.421					.6716	11.5711	-0.7350	0.5402
14	12	17	13.6150	231.4550	289	185.368					.4065	12.2952	1.3198	1.7419
15	13	18	13.5310	243.5580	324	183.088					.5836	13.0193	0.5117	0.2618
16	合计	156	112.7712	836.6494	2054	1083.37					.1183	112.7712	0	9.6928
17	均值	12	8.6747											

18	$t_{0.05/2}(11)$	2.201
19	β1	-0.0145
20	β2	0.7241
21	TSS	105.1183
22	RSS	9.6928
23	ESS	95.4255
24	r2	0.9078
25	sigma_2	0.8812
26	Se_β2	0.0696
27	Se_β1	0.8746
28	t_β2	10.4065
29	t_β1	-0.0165

会看、会算(简要结果报告)
——案例1：受教育程度与时均工资

$$\hat{Y}_i = -0.014\,4 + 0.724\,0X_i$$

se =	(0.931 7)	(0.070 0)
t =	(−0.015 4)	(10.342 8)
p =	(0.987)	(0.000)

$r^2 = 0.906\,5$

df = 11

$F_{1,11} = 108.30$

在原假设H_0:$\beta_i = 0$，下：

（思考）

$$t^*_{\alpha/2} = \frac{\hat{\beta}_i - \beta_i}{S_{\hat{\beta}_i}} = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}$$

会看、会算(Eviews结果报告2)
——案例2 ：英国数据CAPM 模型

OBS	Y	X	OBS	Y	X
1980年1月	6.080 228 52	7.263 448 404	1982年2月	−11.129 075 03	−4.033 607 075
1980年2月	−0.924 185 461	6.339 895 504	1982年3月	1.724 627 956	3.042 525 777
1980年3月	−3.286 174 252	−9.285 216 834	1982年4月	0.157 879 967	0.734 564 665
1980年4月	5.211 976 571	0.793 290 771	1982年5月	−1.875 202 616	2.779 732 288
1980年5月	−16.164 211 11	−2.902 420 985	1982年6月	−10.624 817 67	−5.900 116 576
1980年6月	−1.054 703 649	8.613 150 875	1982年7月	−5.761 135 416	3.005 344 385
1980年7月	11.172 376 99	3.982 062 848	1982年8月	5.481 432 596	3.954 990 619
1980年8月	−11.063 275 51	−1.150 170 907	1982年9月	−17.022 074 59	2.547 127 067
1980年9月	−16.776 996 09	3.486 125 868	1982年10月	7.625 420 708	4.329 008 106
1980年10月	−7.021 834 032	4.329 850 278	1982年11月	−6.575 721 646	0.191 940 594
1980年11月	−9.716 846 68	0.936 875 279	1982年12月	−2.372 829 861	−0.921 675 55
1980年12月	5.215 705 717	−5.202 455 846	1983年1月	17.523 749 36	3.394 682 577
1981年1月	−6.612 000 956	−2.082 757 509	1983年2月	1.354 655 809	0.758 714 353
1981年2月	4.264 498 443	2.728 522 893	1983年3月	16.268 610 49	1.862 073 664
1981年3月	4.916 710 821	0.653 397 106	1983年4月	−6.074 547 158	6.797 751 341
1981年4月	22.204 959 46	6.436 071 962	1983年5月	−0.826 650 702	−1.699 253 628
1981年5月	−11.298 685 24	−4.259 197 932	1983年6月	3.807 881 996	4.092 592 402
1981年6月	−5.770 507 783	0.543 909 707	1983年7月	0.575 700 91	−2.926 299 262
1981年7月	−5.217 764 717	−0.486 845 933	1983年8月	3.755 563 441	1.773 424 306
1981年8月	16.196 201 75	2.843 999 508	1983年9月	−5.365 927 271	−2.800 815 667
1981年9月	−17.169 953 95	−16.457 214 2	1983年10月	−3.750 302 815	−1.505 394 995
1981年10月	1.105 334 728	4.468 938 171	1983年11月	4.898 751 703	4.186 962 84
1981年11月	11.685 336 7	5.885 519 658	1983年12月	4.379 256 151	1.201 416 981
1981年12月	−2.301 451 728	−0.390 698 164	1984年1月	16.560 161 88	6.769 320 788
1982年1月	8.643 728 679	2.499 567 896	1984年2月	1.523 127 464	−1.686 027 417

会看、会算(Eviews结果报告2)
——案例2 ：英国数据CAPM 模型

Dependent Variable: Y
Method: Least Squares
Sample: 1980M01 1999M12
Included observations: 240

	Coefficient	Std. Error	t-Statistic	Prob.
C	−0.447481	0.362943	−1.232924	0.2188
X	1.171128	0.075386	15.53500	0.0000

R-squared	0.503480	Mean dependent var.	0.499826
Adjusted R-squared	0.501394	S.D. dependent var.	7.849594
S.E. of regression	5.542759	Durbin-Watson stat.	1.984746
Sum squared resid.	7311.877	Prob. (F-statistic)	0.000000
F-statistic	241.3363		

可决系数r^2

回归标准误$\hat{\sigma}$

RSS：$\sum e_i^2$

F统计量

因变量均值

因变量标准差

本节内容：

§2.6 双变量线性回归模型的延伸

过原点回归
——资本资产定价模型(CAPM)

- 过原点回归（regression through the origin）
  - 模型没有截距项
  - 在实践中，双变量PRF过原点回归采取如下的形式：

$$Y_i = \beta_2 X_i + u_i \quad （式6.1.1）$$

  - 适用于这种模型的例子：弗里德曼的持久收入假说(permanent income hypothesis)；资本资产定价模型(the capital Asset Pricing Model, CAPM)等。
- 下面以资本资产定价模型(CAPM)为例来加以说明：

$$(ER_i - r_f) = \beta_i (ER_m - r_f) \quad （式6.1.2）$$

其中ER_i为证券i的期望回报率; ER_m为市场证券组合的期望回报率(如标准普尔(S&.P) 500 综合股票指数); r_f为无风险回报率(90 天国债回报率)。β_i为Beta系数，指第i种证券回报率与市场互动程度的度量。

一个大于1的β_i意味着证券i是一种易波动或进攻型证券，而一个小于1 的β_i则意味着证券i是一种防御型证券。(注:不要把这个β_i和双变量回归的斜率系数β_2混同起来。)

过原点回归
——资本资产定价模型(CAPM)

- 如果资本市场能够有效运行，则CAPM要求下式成立：

（式6.1.2）

$$(ER_i - r_f) = \beta_i (ER_m - r_f)$$

证券i 的期望风险溢价　　期望市场风险溢价

- 实证研究常写成下式：

（式6.1.3）

$$R_i - r_f = \beta_i (R_m - r_f) + u_i$$

（式6.1.4）

$$R_i - r_f = \alpha_i + \beta_i (R_m - r_f) + u_i$$

市场模型(Market Model)

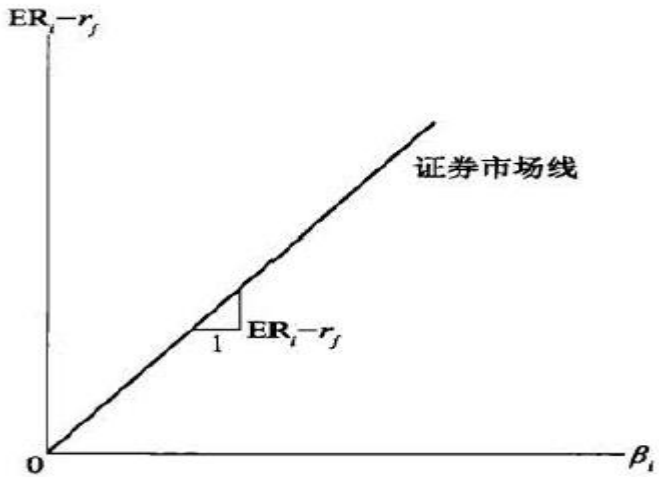


图 6—1　系统风险

①这里的解释变量
为波动性系数β_i，而不是$(ER_m - r_f)$。

②CAPM成立，则预期α_i为零。

这样的模型如何估计呢？

过原点回归
——过原点回归β_2的OLS估计量

- 这类模型的SRF可以写成：

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

- OLS法：$\sum e_i^2 = \sum (Y_i - \hat{\beta}_2 X_i)^2$

$$\frac{d\sum e_i^2}{d\hat{\beta}_2} = 2\sum (Y_i - \hat{\beta}_2 X_i)(-X_i) = 0 \longrightarrow \sum e_i X_i = 0$$

$$\hat{\beta}_2 = \frac{\sum X_i Y_i}{\sum X_i^2} \quad （式6.1.6）$$

$$\hat{\beta}_2 = \frac{\sum X_i(\beta_2 X_i + u_i)}{\sum X_i^2} = \beta_2 + \frac{\sum X_i u_i}{\sum X_i^2} \longleftarrow Y_i = \beta_2 X_i + u_i$$

$$E(\hat{\beta}_2) = \beta_2 \longleftarrow E(\sum X_i u_i) = E(u_i)\sum X_i = 0$$

过原点回归

————过原点回归$var(\hat{\beta}_2)$的OLS估计量和σ^2的无偏估计量

- 这类模型的SRF可以写成：

$$E(\hat{\beta}_2-\beta_2)^2=E[\frac{\sum X_i u_i}{\sum X_i^2}]^2$$

$$\text{var}(\hat{\beta}_2)=\frac{\sigma^2}{\sum X_i^2} \quad \text{（式6.1.7）}$$

$$\hat{\sigma}^2=\frac{\sum \hat{u}_i^2}{n-1},\ \text{E}(\hat{\sigma}^2)=\sigma^2 \quad \text{（式6.1.8）}$$

过原点回归

——OLS估计量对比：无截距和有截距的差异！

（式6.1.2）

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

（式6.1.3）

$$\hat{\beta}_2 = \frac{\sum X_i Y_i}{\sum X_i^2} \quad \longleftrightarrow \quad \hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2}$$

$$\mathrm{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum X_i^2} \quad \longleftrightarrow \quad \mathrm{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2}$$

（式6.1.4）

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-1} \quad \longleftrightarrow \quad \hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$$

第一，对有截距项的模型来说，总有 $\sum e_i = 0$；对无截距项的模型，$\sum e_i = 0$不一定成立，只有 $\sum e_i X_i = 0$。

第二，对有截距项的模型，判定系数$r^2 \geq 0$ ；但是，对无截距模型来说，r^2有时可能出现负值。

过原点回归

——比较：过原点回归的r^2可能小于零！

（式6.1.9）

- 过原点回归的r^2计算公式如下：

$$raw\quad r^2=\frac{\sum(X_iY_i)^2}{\sum X_i^2\sum Y_i^2}$$

$$r^2=1-\frac{RSS}{TSS}=1-\frac{\sum e_i^2}{\sum y_i^2}$$

- 对于有截距模型：

$$RSS=\sum e_i^2=\sum y_i^2-\hat{\beta}_2^2\sum x_i^2\le\sum y_i^2=\text{TSS}$$

$r^2\ge 0$

- 对于无截距模型（过原点回归）：：

$$RSS=\sum e_i^2=\sum Y_i^2-\hat{\beta}_2^2\sum X_i^2$$

可能$\hat{\beta}_2^2\sum X_i^2<N\bar{Y}^2$

$$TSS=\sum y_i^2=\sum Y_i^2-N\bar{Y}^2$$

r^2可能<0

启示：第一，尽管模型含有截距项，但若该项的出现是统计上不显著的(即统计上等于零)，则从任何实际方面考虑，都可认为这个结果是一个过原点回归模型。第二，如果在模型中确实有截距，而我们却执意拟合一个过原点回归，我们就犯了设定错误(specification error)。

过原点回归

——(CAPM)案例说明：104种股票的回报率与总体股市的关系

表 6—1

OBS	Y	X	OBS	Y	X
1980年1月	6.080 228 52	7.263 448 404	1982年2月	−11.129 075 03	−4.033 607 075
1980年2月	−0.924 185 461	6.339 895 504	1982年3月	1.724 627 956	3.042 525 777
1980年3月	−3.286 174 252	−9.285 216 834	1982年4月	0.157 879 967	0.734 564 665
1980年4月	5.211 976 571	0.793 290 771	1982年5月	−1.875 202 616	2.779 732 288
1980年5月	−16.164 211 11	−2.902 420 985	1982年6月	−10.624 817 67	−5.900 116 576
1980年6月	−1.054 703 649	8.613 150 875	1982年7月	−5.761 135 416	3.005 344 385
1980年7月	11.172 376 99	3.982 062 848	1982年8月	5.481 432 596	3.954 990 619
1980年8月	−11.063 275 51	−1.150 170 907	1982年9月	−17.022 074 59	2.547 127 067
1980年9月	−16.776 996 09	3.486 125 868	1982年10月	7.625 420 708	4.329 008 106
1980年10月	−7.021 834 032	4.329 850 278	1982年11月	−6.575 721 646	0.191 940 594
1980年11月	−9.716 846 68	0.936 875 279	1982年12月	−2.372 829 861	−0.921 675 55
1980年12月	5.215 705 717	−5.202 455 846	1983年1月	17.523 749 36	3.394 682 577
1981年1月	−6.612 000 956	−2.082 757 509	1983年2月	1.354 655 809	0.758 714 353
1981年2月	4.264 498 443	2.728 522 893	1983年3月	16.268 610 49	1.862 073 664
1981年3月	4.916 710 821	0.653 397 106	1983年4月	−6.074 547 158	6.797 751 341
1981年4月	22.204 959 46	6.436 071 962	1983年5月	−0.826 650 702	−1.699 253 628
1981年5月	−11.298 685 24	−4.259 197 932	1983年6月	3.807 881 996	4.092 592 402
1981年6月	−5.770 507 783	0.543 909 707	1983年7月	0.575 700 91	−2.926 299 262
1981年7月	−5.217 764 717	−0.486 845 933	1983年8月	3.755 563 441	1.773 424 306
1981年8月	16.196 201 75	2.843 999 508	1983年9月	−5.365 927 271	−2.800 815 667
1981年9月	−17.169 953 95	−16.457 214 2	1983年10月	−3.750 302 815	−1.505 394 995
1981年10月	1.105 334 728	4.468 938 171	1983年11月	4.898 751 703	4.186 962 84
1981年11月	11.685 336 7	5.885 519 658	1983年12月	4.379 256 151	1.201 416 981
1981年12月	−2.301 451 728	−0.390 698 164	1984年1月	16.560 161 88	6.769 320 788
1982年1月	8.643 728 679	2.499 567 896	1984年2月	1.523 127 464	−1.686 027 417

1980-1999 年间104 种股票构成的一个指数的超额回报率Y_t(%)和英国总体股票指数的超额回报率X_t(%)的月度数据，共240 个观测。其中超额回报率指的是超过无风险资产回报率的部分。

$$(ER_i - r_f) = \beta_i (ER_m - r_f)$$

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

过原点回归

——(CAPM)案例说明：104种股票的回报率与总体股市的关系

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

	Coefficient	Std. Error	t-Statistic	Prob.
C	−0.447481	0.362943	−1.232924	0.2188
X	1.171128	0.075386	15.53500	0.0000

R-squared	0.503480	Mean dependent var.	0.499826
Adjusted R-squared	0.501394	S.D. dependent var.	7.849594
S.E. of regression	5.542759	Durbin-Watson stat.	1.984746
Sum squared resid.	7311.877	Prob. (F-statistic)	0.000000
F-statistic	241.3363		

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

注意系数、标准误的差异！

	Coefficient	Std. Error	t-Statistic	Prob.
X	1.155512	0.074396	15.53200	0.0000

R-squared	0.500309	Mean dependent var.	0.499826
Adjusted R-squared†	0.500309	S.D. dependent var.	7.849594
S.E. of regression	5.548786	Durbin-Watson stat.*	1.972853
Sum squared resid.	7358.578		

尺度与测量单位

———案例数据：私人总投资与国内生产总值的关系

- 在回归分析中，因变量Y和解释变量X的测量单位的不同会造成回归结果的差异吗？

表 6—2　　1990—2005 年美国国内私人总投资与 GDP

（除非特别指出，都是以 2000 年美元按链式法则计算；季度数据按季节调整的年率折算）

年份	GPDIBL	GPDIM	GDPB	GDPM
1990	886.6	886 600.0	7 112.5	7 112 500.0
1991	829.1	829 100.0	7 100.5	7 100 500.0
1992	878.3	878 300.0	7 336.6	7 336 600.0
1993	953.5	953 500.0	7 532.7	7 532 700.0
1994	1 042.3	1 042 300.0	7 835.5	7 835 500.0
1995	1 109.6	1 109 600.0	8 031.7	8 031 700.0
1996	1 209.2	1 209 200.0	8 328.9	8 328 900.0
1997	1 320.6	1 320 600.0	8 703.5	8 703 500.0
1998	1 455.0	1 455 000.0	9 066.9	9 066 900.0
1999	1 576.3	1 576 300.0	9 470.3	9 470 300.0
2000	1 679.0	1 679 000.0	9 817.0	9 817 000.0
2001	1 629.4	1 629 400.0	9 890.7	9 890 700.0
2002	1 544.6	1 544 600.0	10 048.8	10 048 800.0
2003	1 596.9	1 596 900.0	10 301.0	10 301 000.0
2004	1 713.9	1 713 900.0	10 703.5	10 703 500.0
2005	1 842.0	1 842 000.0	11 048.6	11 048 600.0

注：GPDIBL＝以 2000 年十亿美元计国内私人总投资。
GPDIM＝以 2000 年百万美元计国内私人总投资。
GDPB＝以 2000 年十亿美元计国内生产总值。
GDPM＝以 2000 年百万美元计国内生产总值。

（Figure 6-2）

尺度与测量单位
———尺度因子的定义

- 对于回归模型：

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \mathrm{e}_i$$

- 尺度因子：ω_1和ω_2分别为Y和X的尺度因子！

（式6.1.7）

$$Y_i^* = \omega_1 Y_i \qquad X_i^* = \omega_2 X_i$$

  - 如果Y_i和X_i是以10亿(billion)美元计量的，我们把它们改为用百万(million)美元去度量，就会有：

（式6.1.8）

$$Y_i^* = 1000 Y_i \qquad X_i^* = 1000 X_i \qquad \omega_1 = \omega_2 = 1000$$

尺度与测量单位
———尺度变换的OLS估计

- 对于模型，并进行数据转换：

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \mathrm{e}_i$$

（式6.2.4）

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

$$Y_i^* = \omega_1 Y_i$$
$$X_i^* = \omega_2 X_i$$
$$e_i^* = \omega_1 e_i$$

- 运用OLS方法估计参数，得到：

（式6.2.11/13）

$$\hat{\beta}_2 * = \frac{\sum x_i^* y_i^*}{\sum x_i^{*2}}$$

$$\mathrm{var}(\hat{\beta}_2^*) = \frac{\sigma^{*2}}{\sum x_i^{*2}}$$

（式6.2.10/12）

$$\hat{\beta}_1^* = \bar{Y}^* - \hat{\beta}_2^* \bar{X}^*$$

$$\mathrm{var}(\hat{\beta}_1^*) = \frac{\sum X_i^{*2}}{n\sum x_i^{*2}} \cdot \sigma^{*2}$$

（式6.2.14）

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

尺度与测量单位

——比较：尺度变换前后OLS估计量的相互关系（数学式）

- 模型对比发现，并进行数据转换：

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \mathrm{e}_i$$

（式6.2.4）

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

$$Y_i^* = \omega_1 Y_i$$

$$X_i^* = \omega_2 X_i$$

$$e_i^* = \omega_1 e_i$$

- OLS估计量有如下关系：

（式6.2.15/19）

$$\hat{\beta}_2^* = \left(\frac{\omega_1}{\omega_2}\right)\hat{\beta}_2$$

$$\mathrm{var}(\hat{\beta}_2^*) = \left(\frac{\omega_1}{\omega_2}\right)^2 \mathrm{var}(\hat{\beta}_2)$$

（式6.2.16/18）

$$\hat{\beta}_1^* = \omega_1 \hat{\beta}_1$$

$$\mathrm{var}(\hat{\beta}_1^*) = \omega_1^2 \,\mathrm{var}(\hat{\beta}_1)$$

（式6.2.17/20）

$$\hat{\sigma}^{*2} = \omega_1^{\,2} \hat{\sigma}^2$$

$$r_{xy}^2 = r_{x^*y^*}^2$$

尺度与测量单位

——比较：尺度变换前后OLS估计量的相互关系（结论）

模型对比，得出如下主要结论：

- (1)当$\omega_1 = \omega_2$，即尺度因子相等时，**斜率**系数及其标准误不受尺度从（Y_i, X_i）到（Y_i^*, X_i^*）的影响。**截距**及其**标准误**却放大或缩小至ω_1倍。
- (2)X尺度不变($\omega_2 = 1$)，Y尺度因子ω_1变化，那么，**斜率**和**截距**系数以及它们各自的**标准误**都要乘以同样的因子ω_1。
- (3)Y尺度不变($\omega_1 = 1$)，而X尺度因子ω_2变化，那么，**斜率**系数及其**标准误**都要乘以因子$\frac{1}{\omega_2}$，而**截距**系数及其标准误不变。

尺度与测量单位

———案例分析：私人总投资与国内生产总值的关系

- GPDI和GDP都以十亿美元计算：

$$\widehat{GPDI}_t = -926.090 + 0.2535\ GDP_t \quad \text{（式6.2.21）}$$

$$se = (116.358)\ (0.0129) \qquad r^2 = 0.9648$$

- GPDI和GDP都以百万美元计算：

$$\widehat{GDPI}_t = -926\ 090 + 0.2535\ GDP_t \quad \text{（式6.2.22）}$$

$$se = (116\ 358)\ (0.0129) \qquad r^2 = 0.9648$$

- GPDI以十亿美元计算而GDP以百万美元计算：

$$\widehat{GPDI}_t = -926.090 + 0.000\ 253\ 5\ GDP_t \quad \text{（式6.2.23）}$$

$$se = (116.358)\ (0.000\ 012\ 9) \qquad r^2 = 0.9648$$

- GPDI以百万美元计算而GDP 以十亿美元计算：

$$\widehat{GPDI}_t = -926\ 090 + 253.524\ GDP_t \quad \text{（式6.2.24）}$$

$$se = (116\ 358.7)\ (12.946\ 5) \qquad r^2 = 0.9648$$

标准化变量回归
———概念及模型设置

- 假设如下双变量回归：

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \mathrm{e}_i$$

- 对Y和X作如下**标准化变换**，得到相应的**标准化变量**：

（式6.3.1/2）

$$Y_i^* = \frac{Y_i - \bar{Y}}{S_Y} \qquad X_i^* = \frac{X_i - \bar{X}}{S_X}$$

标准化变量的特征是：其均值总是0 和标准差总是1。

- 得到如下新的双变量回归模型：

（式6.3.4）

（式6.3.5）

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

对标准化的回归子和回归元做回归，截距项总是零!

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

实际上变成了过原点回归模型!

标准化变量回归
——模型比较与结论

主要结论：

- 第一，由于标准化回归本质上是一个过原点回归，而我们在已经指出通常过原点回归的r^2不能使用，所以我们就没有给出其r^2值。
- 第二，传统模型的β系数与这里的β系数之间存在一种有趣的关系。在双变量情形中，这种关系如下：

（式6.3.8）

$$\hat{\beta}_2^* = \frac{S_X}{S_Y}\hat{\beta}_2$$

证明：自学练习题！

- 第三，在多元回归中，变量标准化可以去除多个自变量之间数量尺度(量纲)的差别，因而具有一定的优点！

标准化变量回归

———案例说明：私人总投资与国内生产总值的关系

- GPDI和GDP都以十亿美元计算：

$$\widehat{GPDI}_t = -926.090 + 0.2535\ GDP_t$$

$$se = (116.358)\ (0.0129) \qquad r^2 = 0.9648$$

若GDP 提高1 美元，则GPDI 平均提高2 5 美分。

（式6.3.1/2）

- 标准化变量后的模型估计：

$$\widehat{GPDI}_t^* = 0.9822\ GDP_t^*$$

$$se = (0.0485)$$

若(标准化) GDP 增加一个标准差，则(标准化) GPDI 平均增加约0. 98 个标准差。

（式6.3.4）

（式6.3.5）

回归模型的函数形式
——三种形式

我们将讨论以下的三种回归模型：

- 对数线性模型
- 半对数模型
- 倒数模型

对数线性模型：学会如何测度弹性
——模型形式

- 指数回归模型（exponential regression model）

（式6.5.1）

$$Y_i = \beta_1 X_i^{\beta_2} e^{u_i}$$

- 可化为：

（式6.5.2）

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

ln表示自然对数

（式6.5.3）

$$\ln Y_i = \alpha + \beta_2 \ln X_i + u_i$$

$$\alpha = \ln \beta_1$$

- 这种模型被称为对数-对数(log-log)，双对数(double-log)或对数一线性(log-linear)模型。进而有：

（式6.5.4）

$$Y_i^* = \alpha + \beta_2 X_i^* + u_i$$

$$Y_i^* = \ln Y_i,\ X_i^* = \ln X_i$$

- 从而可用OLS方法可以得到BLUE估计量：

$$Y_i^* = \hat{\alpha} + \hat{\beta}_2 X_i^* + e_i$$

对数线性模型：学会如何测度弹性

（PRF/SRF）

- 双数线性模型

$$\ln Y_i = \alpha + \beta_2 \ln X_i + u_i \qquad Y_i^* = \hat{\alpha} + \hat{\beta}_2 X_i^* + e_i$$

$$\hat{\alpha} = \ln \hat{\beta}_1$$

- β_2有如下特点：

$$\beta_2 = \frac{d(\ln Y)}{d(\ln X)} = \frac{\frac{1}{Y}dY}{\frac{1}{X}dX} = \frac{dY/Y}{dX/X}$$

斜率β_2就是Y对X的弹性！如果Y代表商品需求量Q，X代表商品价格P，则β_2就表示该商品的需求价格弹性。

- 双数线性模型有如下性质：
  - ①Y对X的弹性在整个研究范围内是常数，一直为β_2，因此这种模型也称为不变弹性模型(constant elasticity model)。
  - ②虽然$\hat{\alpha}$和$\hat{\beta}_2$是无偏估计量，但是进入原始模型的参数β_1的估计值$\hat{\beta}_1$却是有偏估计

$$\beta_1 = anti\log \hat{\alpha}$$

对数线性模型：学会如何测度弹性

（PRF/SRF）

- 双对数线性模型

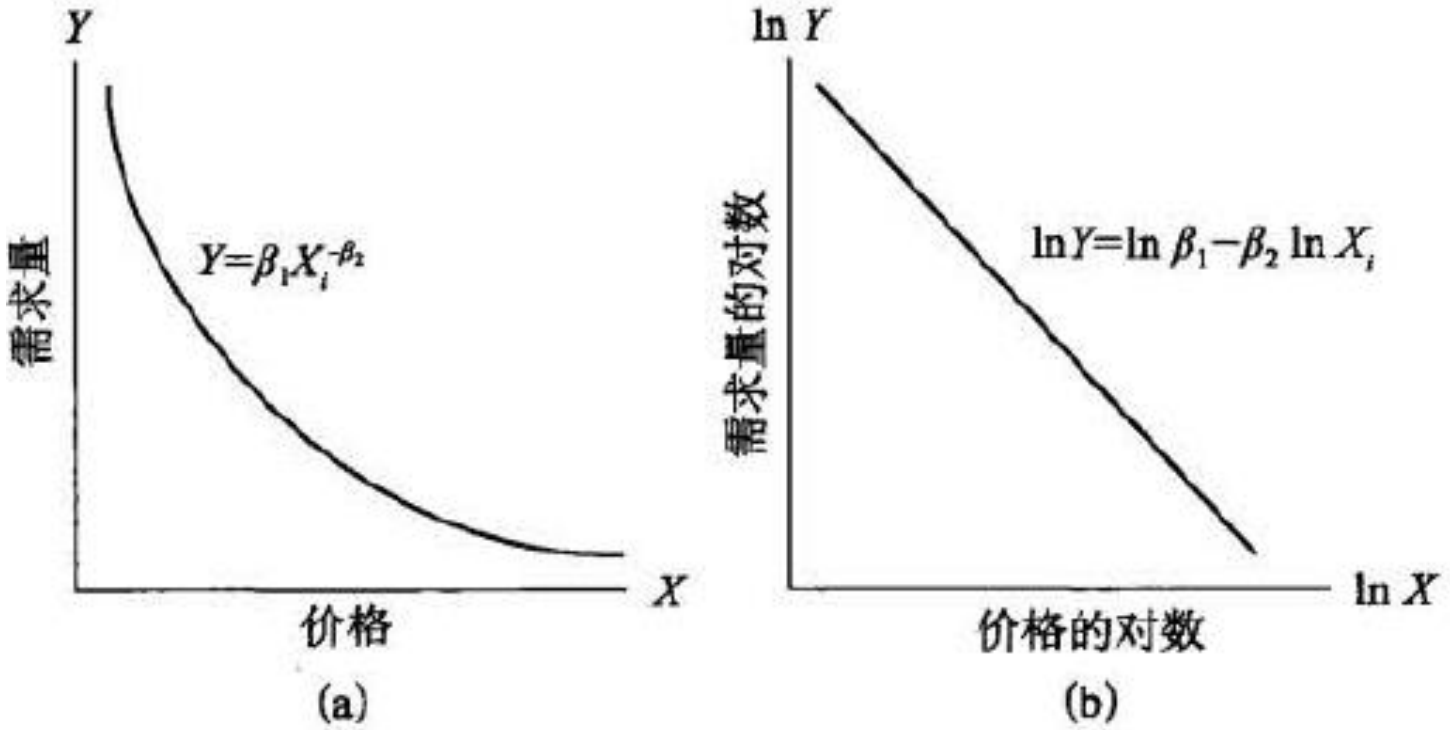


图 6—3　不变弹性模型

对数线性模型：学会如何测度弹性

——案例说明：耐用品支出与个人消费总支出的关系

表 6—3　　个人消费总支出及其分类

（均以 2000 年十亿美元按链式法则计算，季度数据按年增长率进行季节调整）

年份与季度	EXPSERVICES	EXPDUR	EXPNONDUR	PCEXP
2003 - I	4 143.3	971.4	2 072.5	7 184.9
2003 - II	4 161.3	1 009.8	2 084.2	7 249.3
2003 - III	4 190.7	1 049.6	2 123.0	7 352.9
2003 - IV	4 220.2	1 051.4	2 132.5	7 394.3
2004 - I	4 268.2	1 067.0	2 155.3	7 479.8
2004 - II	4 308.4	1 071.4	2 164.3	7 534.4
2004 - III	4 341.5	1 093.9	2 184.0	7 607.1
2004 - IV	4 377.4	1 110.3	2 213.1	7 687.1
2005 - I	4 395.3	1 116.8	2 241.5	7 739.4
2005 - II	4 420.0	1 150.8	2 268.4	7 819.8
2005 - III	4 454.5	1 175.9	2 287.6	7 895.3
2005 - IV	4 476.7	1 137.9	2 309.6	7 910.2
2006 - I	4 494.5	1 190.5	2 342.8	8 003.8
2006 - II	4 535.4	1 190.3	2 351.1	8 055.0
2006 - III	4 566.6	1 208.8	2 360.1	8 111.2

注：EXPSERVICES＝劳务支出，以 2000 年十亿美元为单位。
EXPDUR＝耐用品支出，以 2000 年十亿美元为单位。
EXPNONDUR＝非耐用品支出，以 2000 年十亿美元为单位。
PCEXP＝个人消费总支出，以 2000 年十亿美元为单位。

$$\widehat{\ln \text{EXDUR}_t} = -7.541\,7 + 1.626\,6 \ln \text{PCEX}_t$$

$$se = (0.716\,1) \quad (0.080\,0)$$

$$t = (-10.530\,9)^* \quad (20.315\,2)^* \quad r^2 = 0.969\,5$$

假设我们想求出耐用品支出对个人消费总支出的弹性。将**耐用品支出**的对数相对**个人消费总支出**的对数散点，你将看到二者之间存在线性关系。

因此，双对数模型适用：

半对数模型

——怎样测量增长率：线性到对数模型（The Log-Lin Model）

表 6—3　　个人消费总支出及其分类

（均以 2000 年十亿美元按链式法则计算，季度数据

年份与季度	EXPSERVICES	EXPDUR
2003 - I	4 143.3	971.4
2003 - II	4 161.3	1 009.8
2003 - III	4 190.7	1 049.6
2003 - IV	4 220.2	1 051.4
2004 - I	4 268.2	1 067.0
2004 - II	4 308.4	1 071.4
2004 - III	4 341.5	1 093.9
2004 - IV	4 377.4	1 110.3
2005 - I	4 395.3	1 116.8
2005 - II	4 420.0	1 150.8
2005 - III	4 454.5	1 175.9
2005 - IV	4 476.7	1 137.9
2006 - I	4 494.5	1 190.5
2006 - II	4 535.4	1 190.3
2006 - III	4 566.6	1 208.8

注：EXPSERVICES=劳务支出，以 2000 年十亿美元为单位。
EXPDUR=耐用品支出，以 2000 年十亿美元为单位。
EXPNONDUR=非耐用品支出，以 2000 年十亿美元为单位。
PCEXP=个人消费总支出，以 2000 年十亿美元为单位。

经济学家、企业人员与政府常常对于求出某些经济变量的**增长率**感兴趣，如人口、GNP、货币供给、就业、生产力、贸易赤字等。

（式6.6.1）

$$Y_t = Y_0(1+r)^t$$

Y_t＝时期t的劳务实际支出；
Y_0＝劳务实际支出的初始值(为2002年第四季度末的值);
r是Y的复合增长率

半对数模型：两种形式
——线性到对数模型和对数到线性模型

- 半对数模型的形式：

（式6.6.1）

$$Y_t = Y_0(1+r)^t$$

（式6.6.2）

$$\ln Y_t = \ln Y_0 + t\ln(1+r)$$

（式6.6.5）

$$\ln Y_t = \beta_1 + \beta_2 t$$

$$\beta_1 = \ln Y_0;$$
$$\beta_2 = \ln(1+r)$$

（式6.6.6）

$$\ln Y_t = \beta_1 + \beta_2 t + u_t$$

- 半对数模型(semilog models)：
  - 线性到对数模型(log-lin model)：只有回归子Y取对数
  - 对数到线性模型(lin-log model)：只有回归元X取对数

半对数模型1：线性到对数模型(log-lin model)
——内涵及特征

- **线性到对数模型(log-lin model)**

$$\ln Y_t = \beta_1 + \beta_2 t$$

$$\ln Y_t = \beta_1 + \beta_2 t + u_t$$

$$\beta_1 = \ln Y_0;$$
$$\beta_2 = \ln(1+r)$$

（式6.6.7）

$$\beta_2 = \frac{d\ln Y}{dt} = \frac{dY/Y}{dt} = \frac{\text{因变量的相对改变量}}{\text{自变量的绝对改变量}}$$

半弹性：分子乘以100

- 恒定相对增长率模型：上述模型描述了因变量Y的恒定相对增长率
  - 恒定相对增长模型：$\beta_2 > 0$；
  - 恒定相对衰减模型：$\beta_2 < 0$。

半对数模型1：线性到对数模型(log-lin model)
——劳务支出的相对增长率案例

瞬时增长率

$$\widehat{\ln EXS_t} = 8.322\,6 + 0.007\,05t$$
$$se = (0.001\,6) \quad (0.000\,18) \quad r^2 = 0.991\,9$$
$$t = (5\,201.625)^* \quad (39.166\,7)^*$$

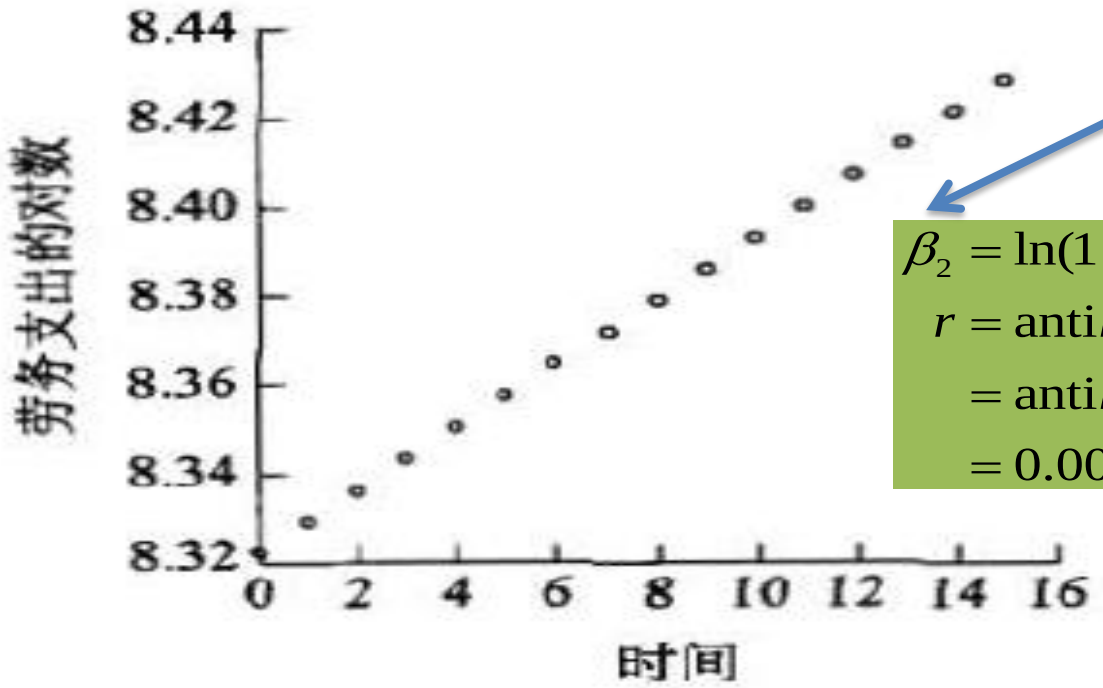


图 6—4

复合增长率

$$\beta_2 = \ln(1+r)$$
$$r = \text{anti}log(\beta_2) - 1$$
$$= \text{anti}log(0.00705) - 1$$
$$= 0.00708$$

（式6.6.8）

半对数模型1：线性到对数模型(log-lin model)
——与线性趋势模型的对比（劳务支出的相对增长率案例）

- 线性趋势模型： Y直接对时间 t回归：

（式6.6.9）

$$Y_t = \beta_1 + \beta_2 t + u_t$$

这里的时间变量t被称为趋势变量（trend variable）。“趋势”是指某种变量中有持续上升或下降的运动。

$$\widehat{EXS}_t = 4\,111.545 + 30.674t$$
$$t = (655.562\,8) \quad (44.467\,1) \qquad r^2 = 0.993\,5$$

解释如下：在2003年第1季度至2006年第3季度期间，劳务支出以每季度约300 亿美元的绝对速度(注意不是相对速度)增加,即劳务支出有上涨的趋势。

半对数模型1：对数到线性模型(lin-log model)
——与线性趋势模型的对比（劳务支出的相对增长率案例）

- **对数到线性模型**(lin-log model)
  - 如果我们的目的是测量X的一个百分比变化时，Y的绝对变化量，则要用对数到线性模型（lin-log model）：

（式6.6.11）

$$Y_t = \beta_1 + \beta_2 \ln X_i + u_t$$

（式6.6.12）

$$\beta_2 = \frac{Y\text{的绝对变化}}{X\text{的相对变化}} = \frac{dY}{d\ln X} = \frac{dY}{dX/X} = \frac{\Delta Y}{\Delta X/X}$$

（式6.6.13）

$$\Delta Y = \beta_2 \frac{\Delta X}{X}$$

  - 例如：恩格尔支出(Engel expenditure) 模型：
    - “用于食物的总支出以算术级数增加，而总支出以几何级数增加。”

半对数模型1：对数到线性模型(lin-log model)

——案例说明：印度55个家庭的食物支出和总支出的关系

表 2—8　　食物支出与总支出　　（单位：卢比）

观测	食物支出	总支出	观测	食物支出	总支出
1	217.000 0	382.000 0	29	390.000 0	655.000 0
2	196.000 0	388.000 0	30	385.000 0	662.000 0
3	303.000 0	391.000 0	31	470.000 0	663.000 0
4	270.000 0	415.000 0	32	322.000 0	677.000 0
5	325.000 0	456.000 0	33	540.000 0	680.000 0
6	260.000 0	460.000 0	34	433.000 0	690.000 0
7	300.000 0	472.000 0	35	295.000 0	695.000 0
8	325.000 0	478.000 0	36	340.000 0	695.000 0
9	336.000 0	494.000 0	37	500.000 0	695.000 0
10	345.000 0	516.000 0	38	450.000 0	720.000 0
11	325.000 0	525.000 0	39	415.000 0	721.000 0
12	362.000 0	554.000 0	40	540.000 0	730.000 0
13	315.000 0	575.000 0	41	360.000 0	731.000 0
14	355.000 0	579.000 0	42	450.000 0	733.000 0
15	325.000 0	585.000 0	43	395.000 0	745.000 0
16	370.000 0	586.000 0	44	430.000 0	751.000 0
17	390.000 0	590.000 0	45	332.000 0	752.000 0
18	420.000 0	608.000 0	46	397.000 0	752.000 0
19	410.000 0	610.000 0	47	446.000 0	769.000 0
20	383.000 0	616.000 0	48	480.000 0	773.000 0
21	315.000 0	618.000 0	49	352.000 0	773.000 0
22	267.000 0	623.000 0	50	410.000 0	775.000 0
23	420.000 0	627.000 0	51	380.000 0	785.000 0
24	300.000 0	630.000 0	52	610.000 0	788.000 0
25	410.000 0	635.000 0	53	530.000 0	790.000 0
26	220.000 0	640.000 0	54	360.000 0	795.000 0
27	403.000 0	648.000 0	55	305.000 0	801.000 0
28	350.000 0	650.000 0			

印度的55 个农户的食物支出和总支出数据

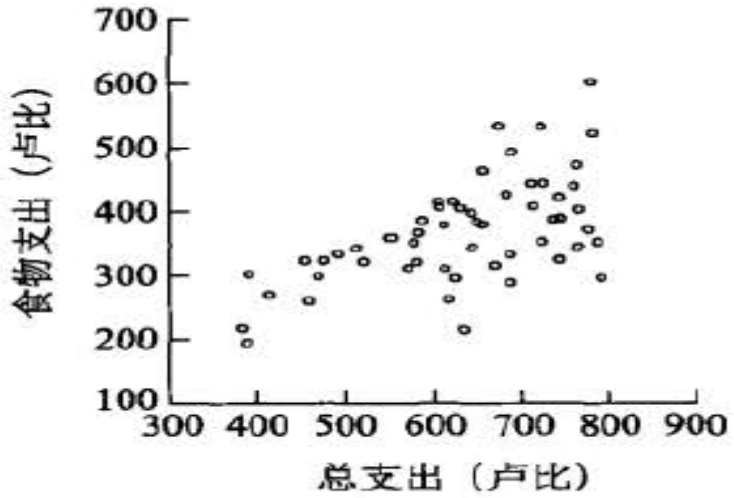


图 6—5

$$\widehat{\text{FoodExp}}_i = -1\,283.912 + 257.270\,0\ \ln\text{TotalExp}_i$$

$$t = (-4.384\,8)^{*}\ (5.662\,5)^{*} \qquad r^2 = 0.376\,9$$

解释：总支出每提高1%,导致家庭的食物支出平均增加约2.57卢比.

倒数模型
——形式、特征及应用

- **倒数模型（Reciprocal Model）：**

$$Y_i = \beta_1 + \beta_2(\frac{1}{X_i}) + u_i \quad （式6.7.1）$$

它是一个线性模型吗？

a.平均固定成本(AFC)曲线

b.菲利普斯曲线（Phillips curve）

c.恩格尔曲线（the Engel expenditure curve）

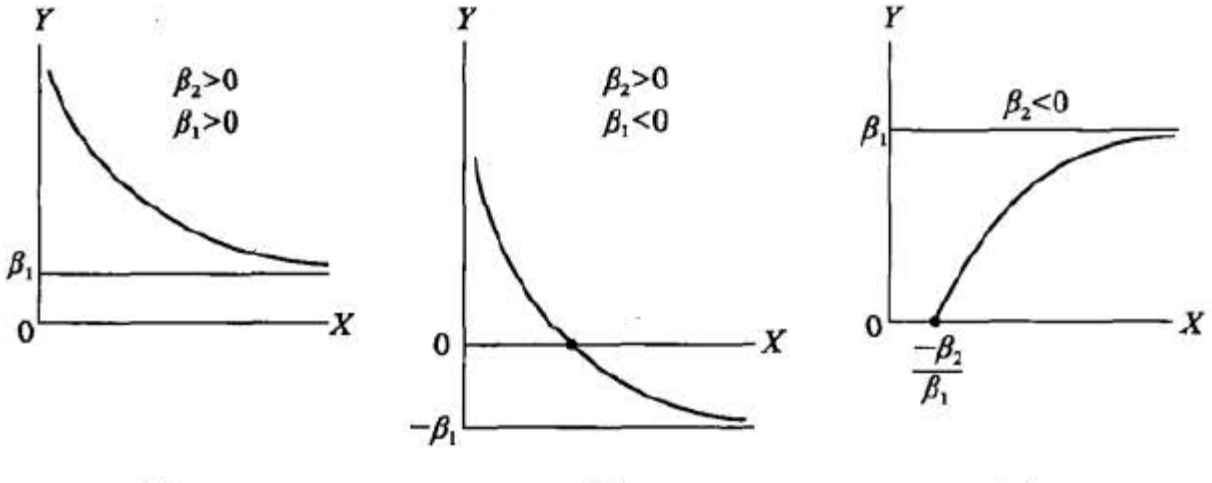


图 6—6 倒数模型：$Y=\beta_1+\beta_2\left(\frac{1}{X}\right)$

- 特征：总有一条内在的渐近线！

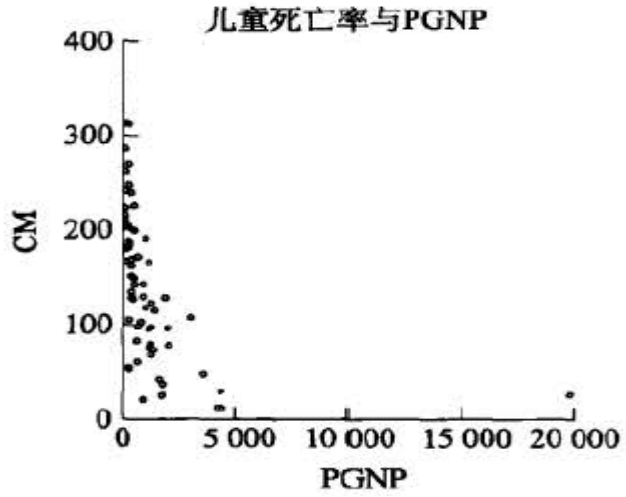
$$X \to \infty \Rightarrow \beta_2(\frac{1}{X}) \to 0 \Rightarrow Y \to \beta_1$$

倒数模型
——说明案例a：儿童死亡率与人均GNP的关系

表 6—4　　64 个国家的生育率及其他数据

观测	CM	FLR	PGNP	TFR	观测	CM	FLR	PGNP	TFR
1	128	37	1 870	6.66	8	240	29	300	5.89
2	204	22	130	6.15	9	241	11	120	5.89
3	202	16	310	7.00	10	55	55	290	2.36
4	197	65	570	6.25	11	75	87	1 180	3.93
5	96	76	2 050	3.81	12	129	55	900	5.99
6	209	26	200	6.44	13	24	93	1 730	3.50
7	170	45	670	6.19	14	165	31	1 150	7.41

注：CM=儿童死亡率，每千名儿童中每年不足 5 岁便死亡的儿童人数。
FLR=妇女识字率，%。
PGNP=1980 年的人均 GNP。



$$\widehat{CM}_i = 81.794\,36 + 27\,237.17\,(1/PGNP_i)$$

$$se = (10.832\,1)\quad(3\,759.999)$$

$$t = (7.551\,1)\quad(7.253\,5)\qquad r^2 = 0.459\,0$$

倒数模型
——说明案例b：菲利普斯曲线（Phillips curve）

- 菲利普斯曲线（Phillips curve）

  工资变化对失业水平的反应中，存在不对称性：
  - 当失业率低于经济学家所称的自然失业率U^N时，由失业的单位变化引起的工资上升要快于当失业率高于自然水平时由失业的同样变化引起的工资下降。
  - β_1表示工资变化的渐近底限。菲利普斯曲线的这一具体特征可能缘于工会的讨价还价能力、最低工资规定和失业补贴等制度因素。

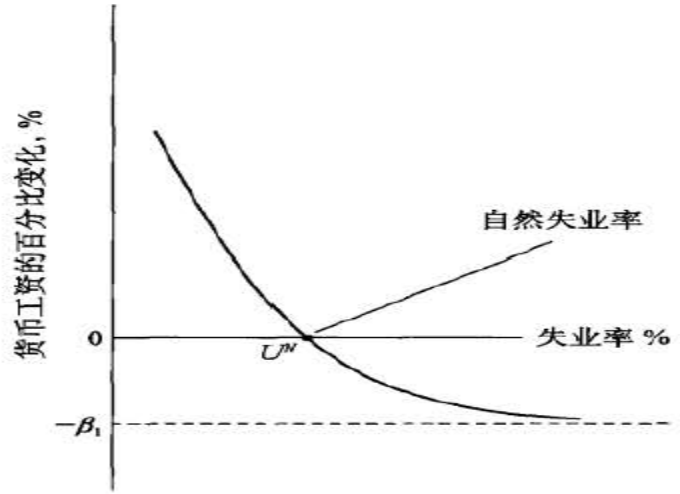


图 6—8　菲利普斯曲线

倒数模型
——说明案例b：菲利普斯曲线（Phillips curve）

表 6—5　美国通货膨胀率与失业率：1960—2006 年

（对所有城市消费者；除非特别指出，否则令 1982—1984=100）

观测	通货膨胀率	失业率	观测	通货膨胀率	失业率
1960	1.718	5.5	1984	4.317	7.5
1961	1.014	6.7	1985	3.561	7.2
1962	1.003	5.5	1986	1.859	7.0
1963	1.325	5.7	1987	3.650	6.2
1964	1.307	5.2	1988	4.137	5.5
1965	1.613	4.5	1989	4.818	5.3
1966	2.857	3.8	1990	5.403	5.6
1967	3.086	3.8	1991	4.208	6.8
1968	4.192	3.6	1992	3.010	7.5
1969	5.460	3.5	1993	2.994	6.9
1970	5.722	4.9	1994	2.561	6.1
1971	4.381	5.9	1995	2.834	5.6
1972	3.210	5.6	1996	2.953	5.4
1973	6.220	4.9	1997	2.294	4.9
1974	11.036	5.6	1998	1.558	4.5
1975	9.12				
1976	5.76				
1977	6.50				
1978	7.59				
1979	11.35				
1980	13.499	7.1	2004	2.663	5.5
1981	10.316	7.6	2005	3.388	5.1
1982	6.161	9.7	2006	3.226	4.6
1983	3.212	9.6			

注：通货膨胀率为 CPI 的年百分比变化。失业率为城市失业率。

即使失业率无限增加，通货膨胀率的最大变化也就是下降约3.07个百分点。

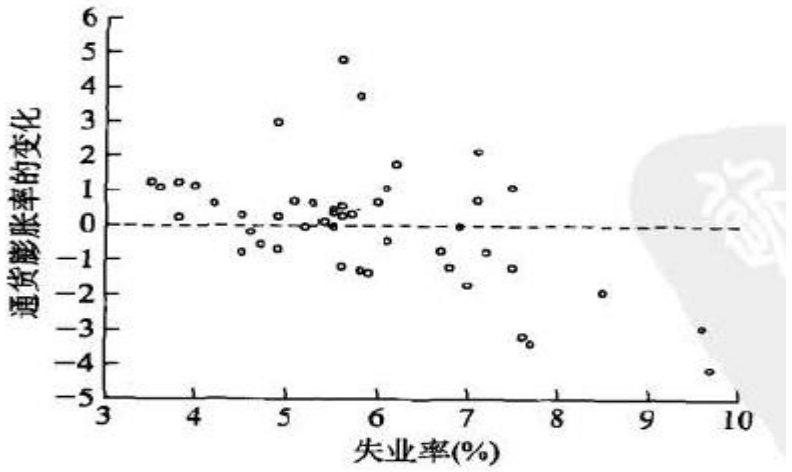


图 6—9　修正的菲利普斯曲线

线性模型：$\widehat{(\pi_t - \pi_{t-1})} = 3.7844 - 0.6385 UN_t$

$t = (4.1912)\ (-4.2756) \quad r^2 = 0.2935$

倒数模型：$\widehat{(\pi_t - \pi_{t-1})} = -3.0684 + 17.2077 \left(\frac{1}{UN_t}\right)$

$t = (-3.1635)\ (3.2886) \quad r^2 = 0.1973$

$$U^N = \frac{\hat{\beta}_1}{-\hat{\beta}_2} = \frac{3.7844}{0.6385} = 5.9270$$

函数形式的选择
——一些技巧和经验

- **选择适当模型时，需要一些技巧和经验：**
  - 1. 模型背后的理论(如菲利普斯曲线〉可能给出了一个特定的函数形式。
  - 2. 最好能求出回归子相对回归元的变化率〈即斜率〉和回归子对回归元的弹性(见下页ppt)。
  - 3. 所选模型的系数应该满足一定的先验预期。
  - 4. 有时多个模型都能相当不错地拟合一个给定的数据集。
  - 5. 通常不应该过分强调r^2这个指标
  - 6. 在有些情形中，确定一个特定的函数形式不是那么容易，此时我们或许可以使用所谓的博克斯-考克斯变换(Box-Cox transformations)

函数形式的选择
——一些供参考的函数形式

表 6—6

模型	方程	斜率 $\left(=\frac{dY}{dX}\right)$	弹性 $\left(=\frac{dY}{dX}\cdot\frac{X}{Y}\right)$
线性	$Y=\beta_1+\beta_2 X$	β_2	$\beta_2\left(\frac{X}{Y}\right)^*$
对数线性 (对数—对数)	$\ln Y=\beta_1+\beta_2\ln X$	$\beta_2\left(\frac{Y}{X}\right)$	β_2
线性到对数	$\ln Y=\beta_1+\beta_2 X$	$\beta_2(Y)$	$\beta_2(X)^*$
对数到线性	$Y=\beta_1+\beta_2\ln X$	$\beta_2\left(\frac{1}{X}\right)$	$\beta_2\left(\frac{1}{Y}\right)^*$
倒数	$Y=\beta_1+\beta_2\left(\frac{1}{X}\right)$	$-\beta_2\left(\frac{1}{X^2}\right)$	$-\beta_2\left(\frac{1}{XY}\right)^*$
对数倒数	$\ln Y=\beta_1-\beta_2\left(\frac{1}{X}\right)$	$\beta_2\left(\frac{Y}{X^2}\right)$	$\beta_2\left(\frac{1}{X}\right)^*$

注：*表示弹性系数是可变的，它依赖于 X 或 Y 或二者的取值。在 X 和 Y 未给定时，实践中常常在均值 $\bar{X}$ 和 $\bar{Y}$ 处测度这些弹性。