计量经济学(Econometrics)

胡华平

西北农林科技大学

经济管理学院数量经济教研室

huhuaping01@hotmail.com

2023-02-15

第2章：一元回归的基本思想

2.1 “回归”的历史渊源

2.2 术语与符号

2.3 数据的类型和性质

2.4 一个假想的微型总体

2.5 一些重要的概念

2.6 总体回归

2.7 样本回归

2.1 “回归”的历史渊源

身高数据

皮尔逊和高尔顿的身高数据（1888年）：

Heights of the Mid-parents in inches. 父辈身高	Heights of the Adult Children.														Total Number of 子辈人数 父辈人数		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	子辈身高
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

回归到中等

高尔顿的发现：

- 父母高，儿女也高；父母矮，儿女也矮
- 但给定父母的身高，儿女的平均身高却趋向于或者"回归"到全体人口的平均身高。

皮尔逊的证实：

- 收集一些家庭群体的一千多名成员的身高记录。
- 两组样本：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。

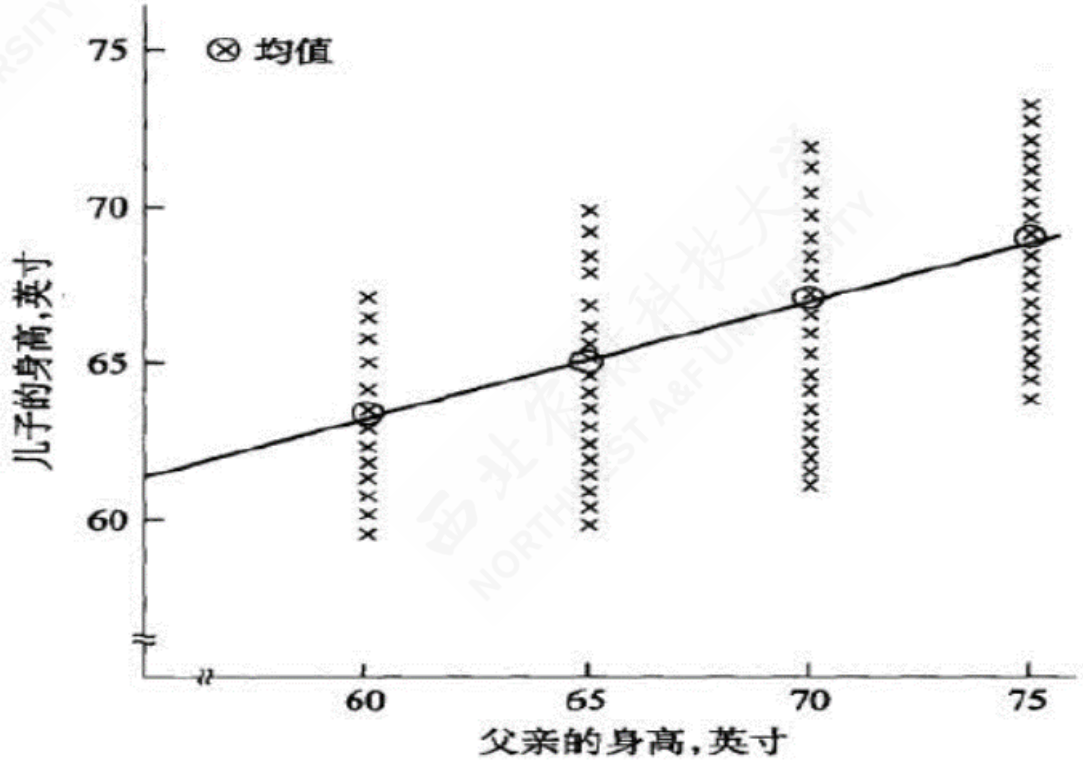
“回归到中等”(regression to mediocrity)的趋势，回归由此而得名

“回归”的现代解释

- 简单地，回归分析研究被解释变量（Y）对一个或多个解释变量（X）之间的依赖关系及规律性。
- 正式地，回归分析通过解释变量（抽样样本中）的观测值（X），去估计和（或）预测被解释变量（Y）的**均值**（总体期望）。

案例说明：子辈身高

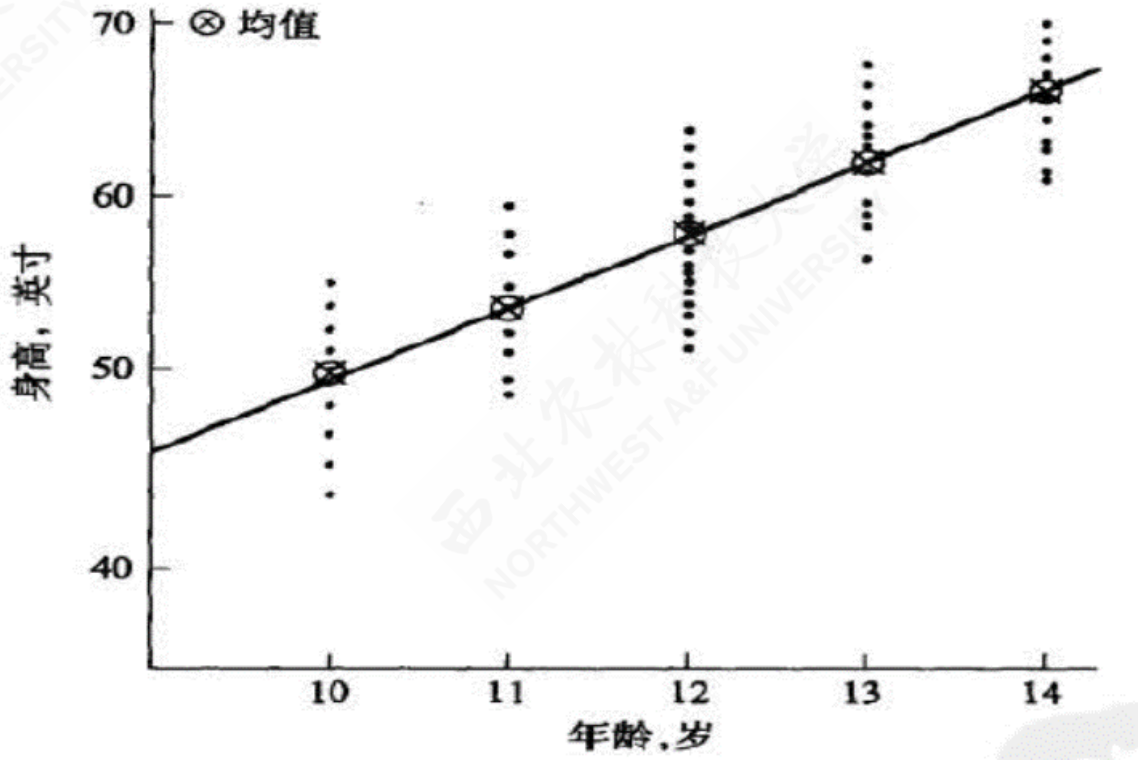
- 给定父亲身高，在一个假想人口总体中的子辈身高分布。



给定父亲身高时儿子身高的假想分布

案例说明：年龄身高

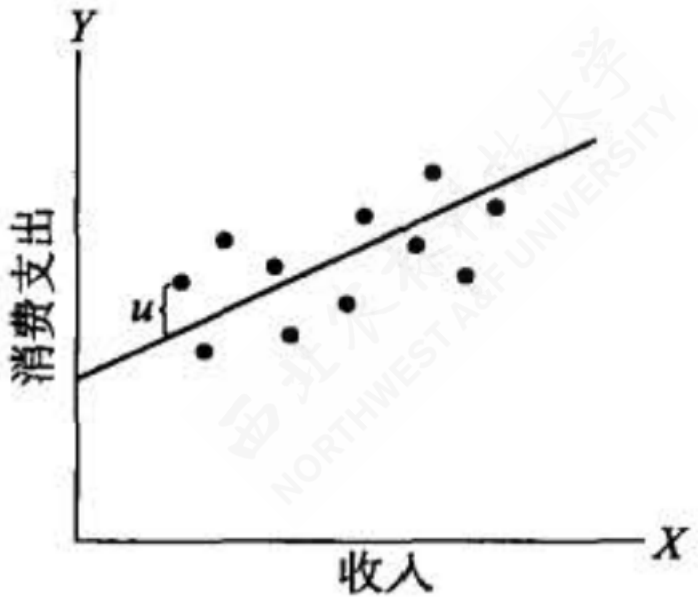
- 给定年龄，男孩子身高总体的分布。



对应于选定年龄的假想身高分布

案例说明：消费函数

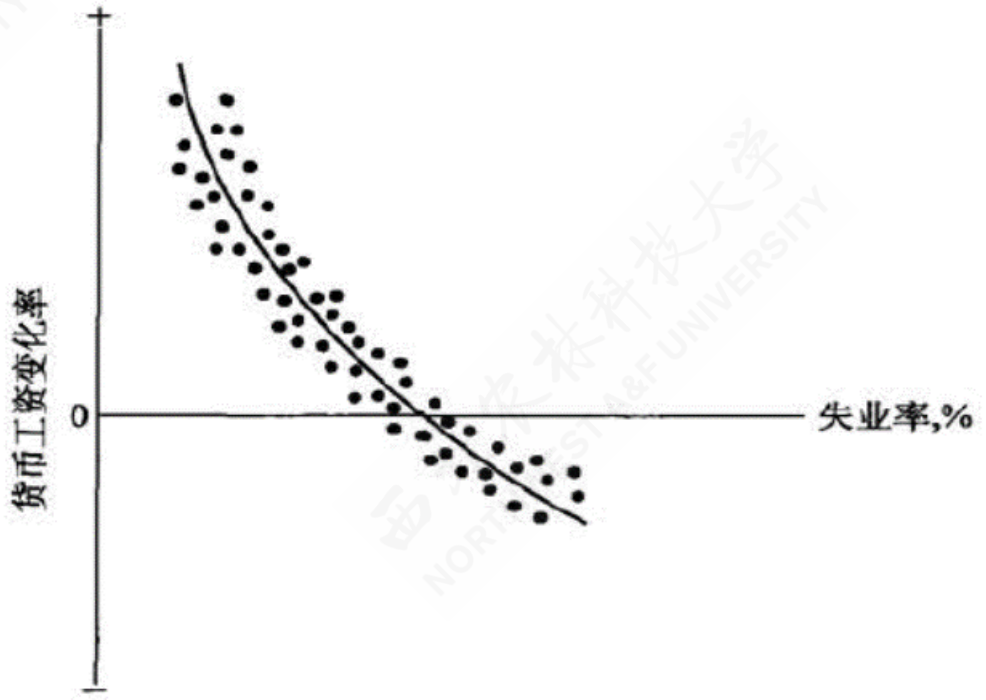
给定税后或可支配收入，个人消费是如何分布的。这种分析有助于估计边际消费倾向（MPC），就是实际收入每美元价值的变化所引起的消费支出的平均变化。



线性形式的凯恩斯消费模型

案例说明：货币工资

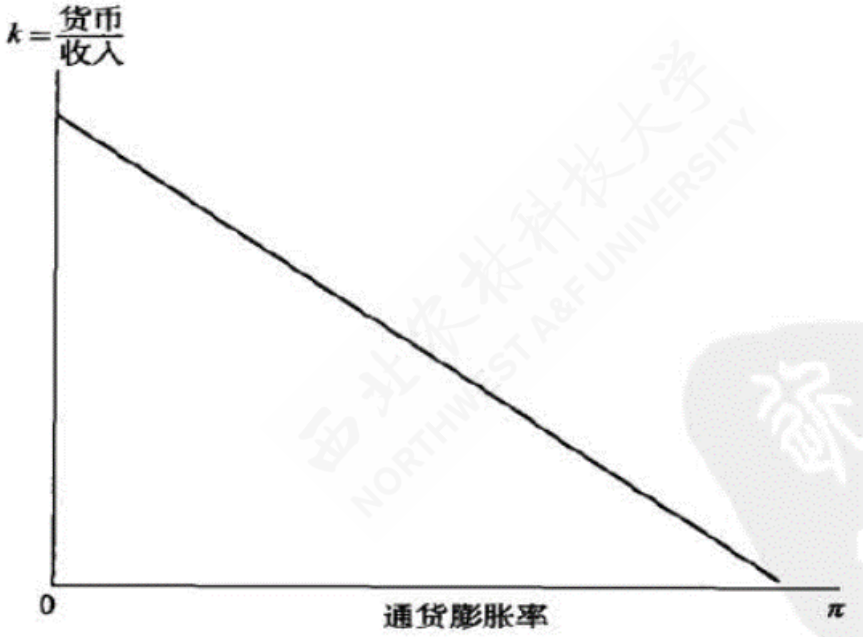
失业率是怎样影响货币工资变化的。



假想的菲利普斯曲线

案例说明：货币持有

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。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



货币持有与通货膨胀率的关系

2.2 术语与符号

X和Y的各种术语约定

Y	X
被解释变量(Explained variable)	解释变量(Explanatory variable)
因变量(Dependent variable)	自变量(Independent 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)	外生(Exogenous)
结果变量(Outcome)	协变量(Covariate)
被控变量(Controlled variable)	控制变量(Control variable)

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

多元回归分析(multiple regression analysis)：

- 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；

模型符号

- 因变量：Y，具体记为 Y_i,
- 解释变量：X，记为 $X_1, X_2, \cdots, X_k$.
  - X_k具体记为：$X_{k1}, X_{k2}, \cdots, X_{kn}$.
  - X_k代表第 k个解释变量，下标 i（或 t）则表示第 i（或 t）个观测值。

情景符号

- 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）

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。

两套体系

李子奈的k元回归：

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + u_i$$

古扎拉蒂的k变量回归：

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

2.3 数据的类型和性质

Type1：时间序列数据(time series data)

时间序列数据：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

Type1：时间序列数据(time series data)：

平稳性(stationary)：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)。

1951年1月-1999年9月美国的M1货币供给

Type2：截面数据(cross-section data)

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

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。

案例：鸡蛋价格与鸡蛋产量

美国50个州的蛋类生产和价格数据

STATE	Y1	X1
AL	2206	92.7
AK	0.7	151
AZ	73	61
AR	3620	86.3
CA	7472	63.4

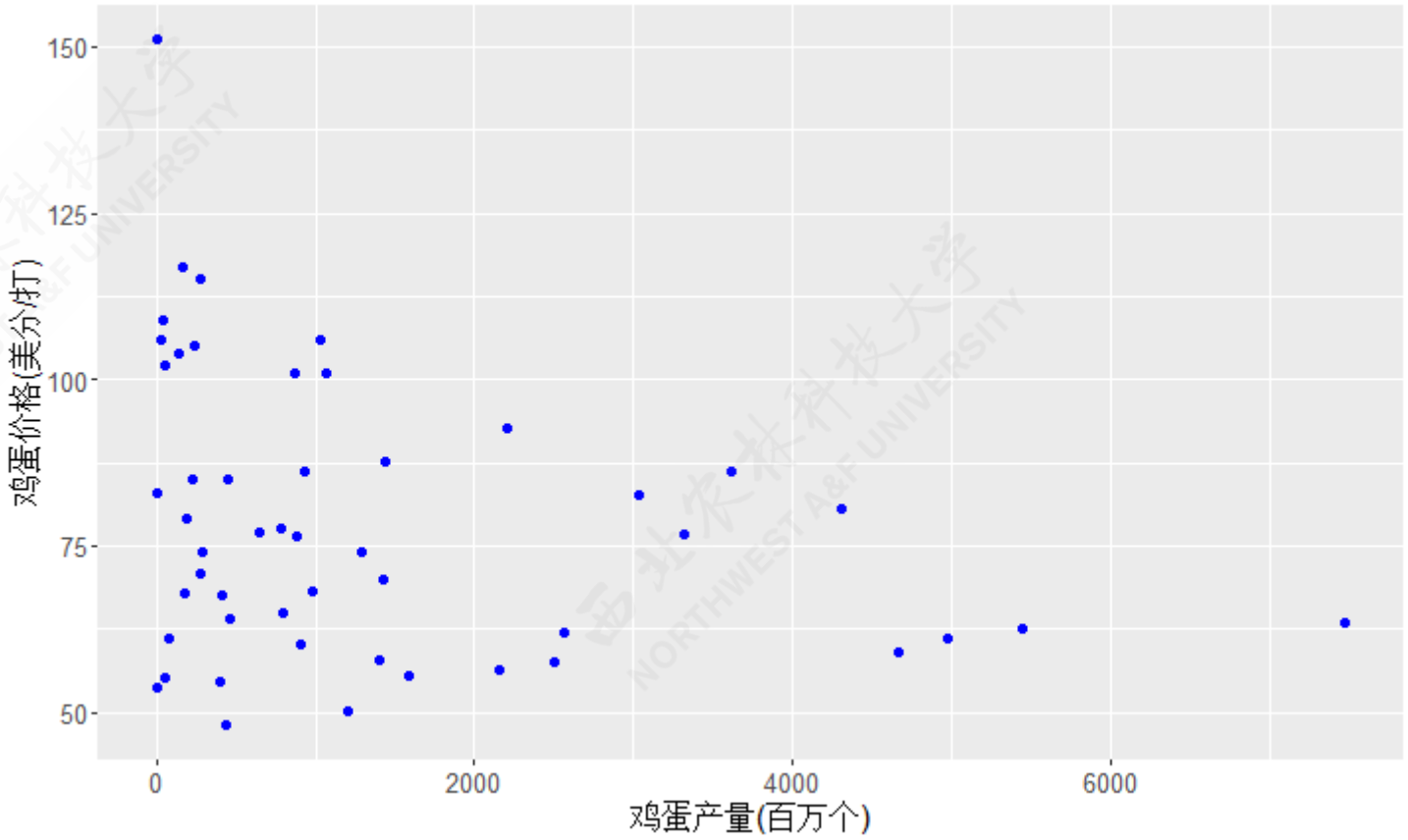
Showing 1 to 5 of 50 entries

Previous 1 2 3 4 5 … 10 Next

美国50个州的蛋类生产和价格数据，其中：

- Y_1代表1990年鸡蛋产量(百万个)；
- X_1代表1990年每打鸡蛋的价格(美分/打)。

案例：鸡蛋价格与鸡蛋产量



1990年蛋产量与价格的关系

思考提问：图中特征符合经济学理论么？图中反映了数据可能存在哪些潜在问题？

Type3：面板数据(Panel Data)

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） n：

- 数据点（观测数）=截面单元数时期数 $n = q * t$

可能存在的问题：

- “平稳性”问题：
- “异方差”问题：

案例：钢铁公司

两家钢铁公式的数据案例：

- 公司：GE=通用公司；US=美国钢铁
- I=真实总投资（百万美元）
- F=前一年的企业真实价值（百万美元）
- C=前一年的真实资本存量（百万美元）

案例：钢铁公司

扁数据形式：

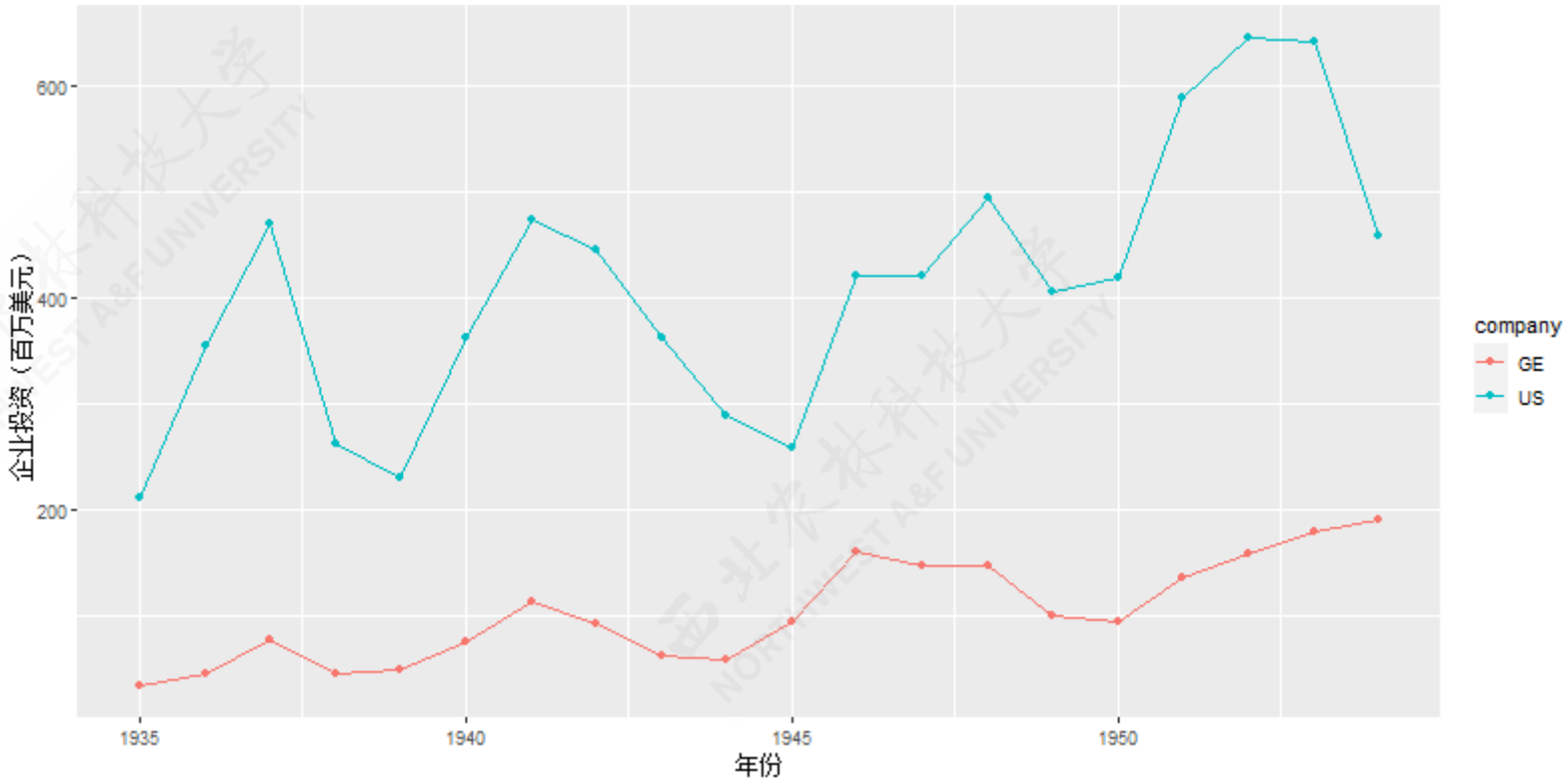
1935-1954年间美国两大钢铁公司的数据(扁数据)

year	GE.C	GE.F	GE.I	US.C	US.F	US.I
1935	97.8	1170.6	33.1	53.8	1362.4	209.9
1936	104.4	2015.8	45	50.5	1807.1	355.3
1937	118	2803.3	77.2	118.1	2673.3	469.9
1938	156.2	2039.7	44.6	260.2	1801.9	262.3
1939	172.6	2256.2	48.1	312.7	1957.3	230.4
1940	186.6	2132.2	74.4	254.2	2202.9	361.6
1941	220.9	1834.1	113	261.4	2380.5	472.8
1942	287.8	1588	91.9	298.7	2168.6	445.6

Showing 1 to 8 of 20 entries

Previous 1 2 3 Next

案例：钢铁公司



两家公司的企业投资情况

案例：钢铁公司

长数据形式：

1935-1954年间美国两大钢铁公司的数据(长数据)

year	company	I	F	C
1935	GE	33.1	1170.6	97.8
1936	GE	45	2015.8	104.4
1937	GE	77.2	2803.3	118
1938	GE	44.6	2039.7	156.2
1939	GE	48.1	2256.2	172.6
1940	GE	74.4	2132.2	186.6
1941	GE	113	1834.1	220.9

Showing 1 to 7 of 40 entries

Previous 1 2 3 4 5 6 Next

案例：钢铁公司

缺失部分数据：

1935-1954年间美国两大钢铁公司的数据(缺失部分数据)

year	GE.C	GE.F	GE.I	US.C	US.F	US.I
1935	97.8	1170.6	33.1	53.8	1362.4	209.9
1936	104.4	2015.8	45	50.5	1807.1	355.3
1937	118	2803.3	77.2	118.1	2673.3	469.9
1938	156.2	2039.7	44.6	260.2	1801.9	262.3
1939	172.6	2256.2	48.1	312.7	1957.3	230.4
1940	186.6	2132.2	74.4			
1941	220.9	1834.1	113	261.4	2380.5	472.8

Showing 1 to 7 of 20 entries

Previous 1 2 3 Next

课堂测试：问1：平衡面板还是非平衡面板？问2：多少数据点？

数据的性质和层次

数据不是“平等”的，也有“三六九等”：

- 名义尺度(nominal scale)
- 序数尺度(ordinal scale)
- 区间尺度(interval scale)
- 比率尺度(ratio scale)

名义尺度(nominal scale)

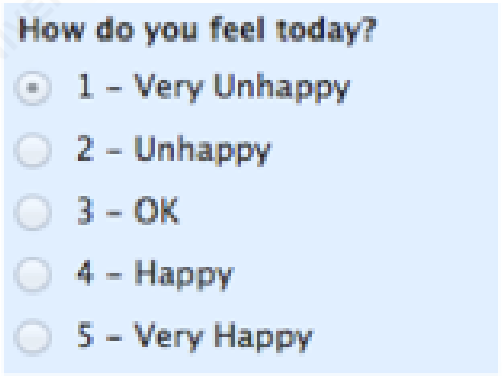
名义尺度变量只表示不同的类别，它不能加减乘除，也不能比较大小。

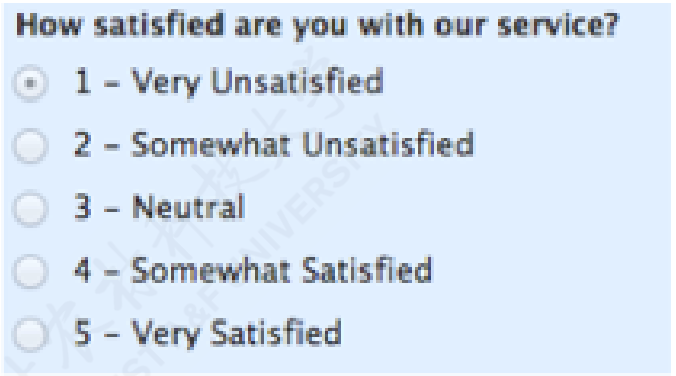
- 如性别(男、女）和婚姻状况（已婚、未婚、离婚、分居）之类的变量。

序数尺度(ordinal scale)

名义尺度变量只能比较大小(即自然顺序)，不能加减乘除。

- 五分量表

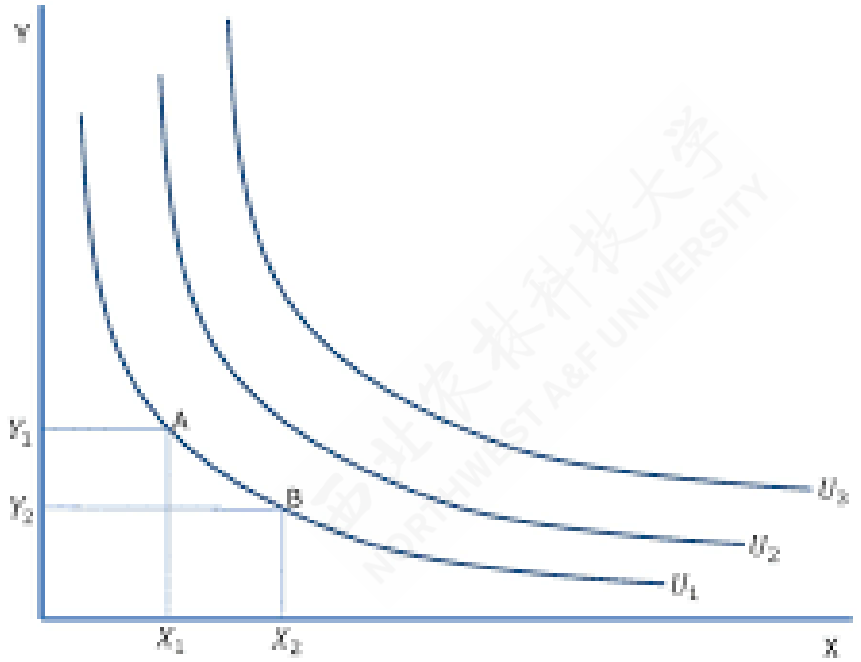




李克特量表

序数尺度(ordinal scale)

- 无差异曲线



两种商品消费下的无差异曲线

区间尺度(interval scale)

区间尺度变量比率尺度变量可以比较大小，也能加减，但不能乘除。

- 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。

比率尺度(ratio scale)：

比率尺度变量可以比较大小，也能加减乘除。

- 对于一个变量X，取其两个值 X_1和 X_2，比率 X_1/X_2和距离 $(X_2 - X_1)$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）(性质3)。因此如 $X_2 \leq X_1$或 $X_2 \geq X_1$之类的比较也是有意义的。
- 如：GDP(亿元)、个人收入(元)等

2.4 一个假想的微型世界

60个家庭的微型总体数据

直观列表：

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
小计	325	462	445	707	678	750	685	1043	966	1211
合计	**7272**									

60个家庭的微型总体数据

扁数据形态："非标准"数据形态（但很直观）

60个家庭的收入和支出情况：假设的总体

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

Showing 1 to 8 of 8 entries

Previous 1 Next

60个家庭的微型总体数据

长数据形态：标准数据形态（但不直观）

60个家庭的收入和支出情况：假设的总体

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

Showing 1 to 7 of 60 entries

Previous 1 2 3 4 5 … 9 Next

2.5 一些重要概念

无条件概率和无条件期望

无条件概率：

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

无条件期望：

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

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

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

无条件概率和无条件期望的示例计算

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

无条件概率和无条件期望

无条件期望的计算过程

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

条件概率和条件期望

条件概率：

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

条件期望：

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

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

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

条件概率和条件期望的示例计算

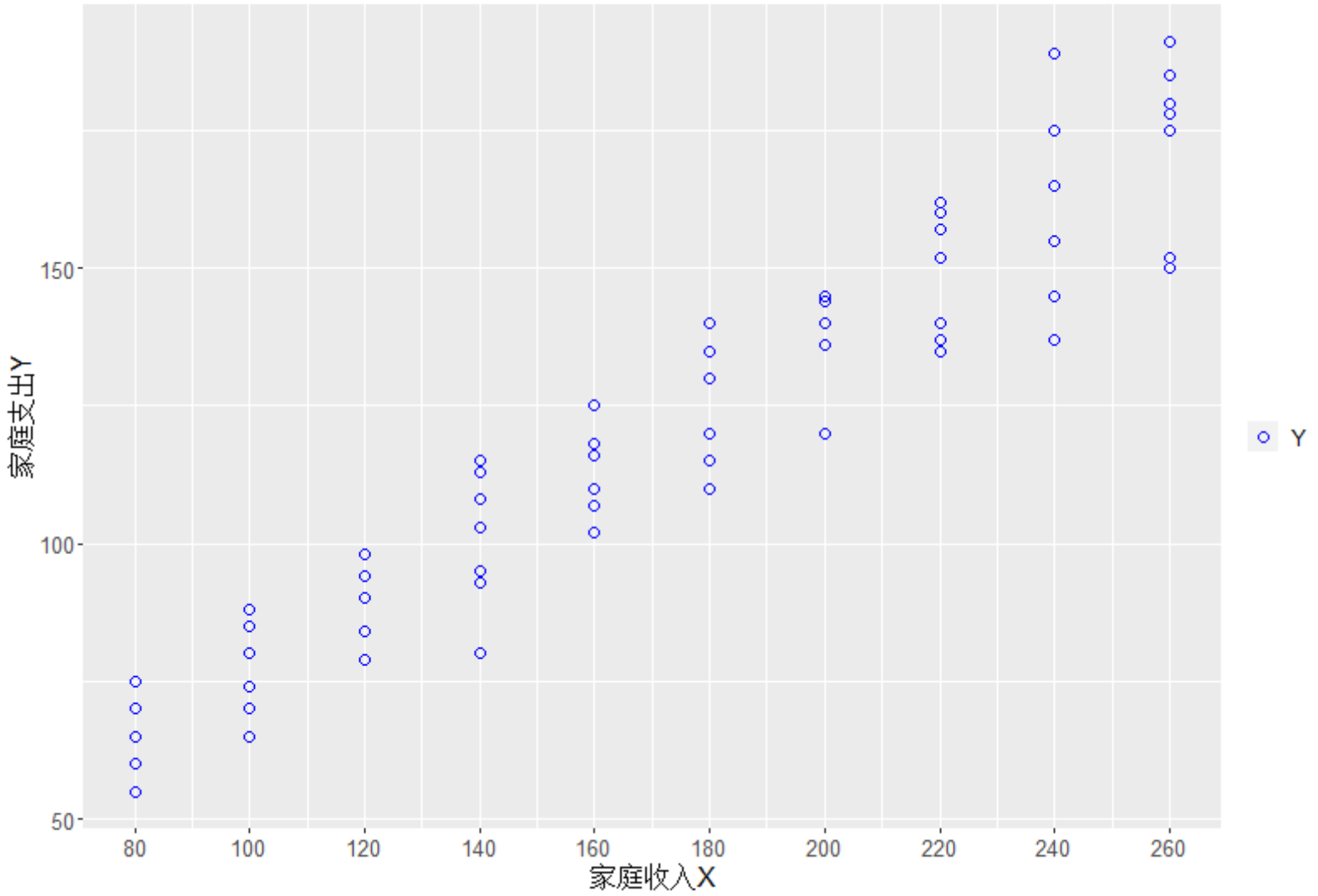
Y，每周家庭消费支出	X，每周家庭收入（美元）																			
	80		100		120		140		160		180		200		220		240		260	
	55	1/5	65	1/6	79	1/5	80	1/7	102	1/6	110	1/6	120	1/5	135	1/7	137	1/6	150	1/7
	60	1/5	70	1/6	84	1/5	93	1/7	107	1/6	115	1/6	136	1/5	137	1/7	145	1/6	152	1/7
	65	1/5	74	1/6	90	1/5	95	1/7	110	1/6	120	1/6	140	1/5	140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

条件概率和条件期望

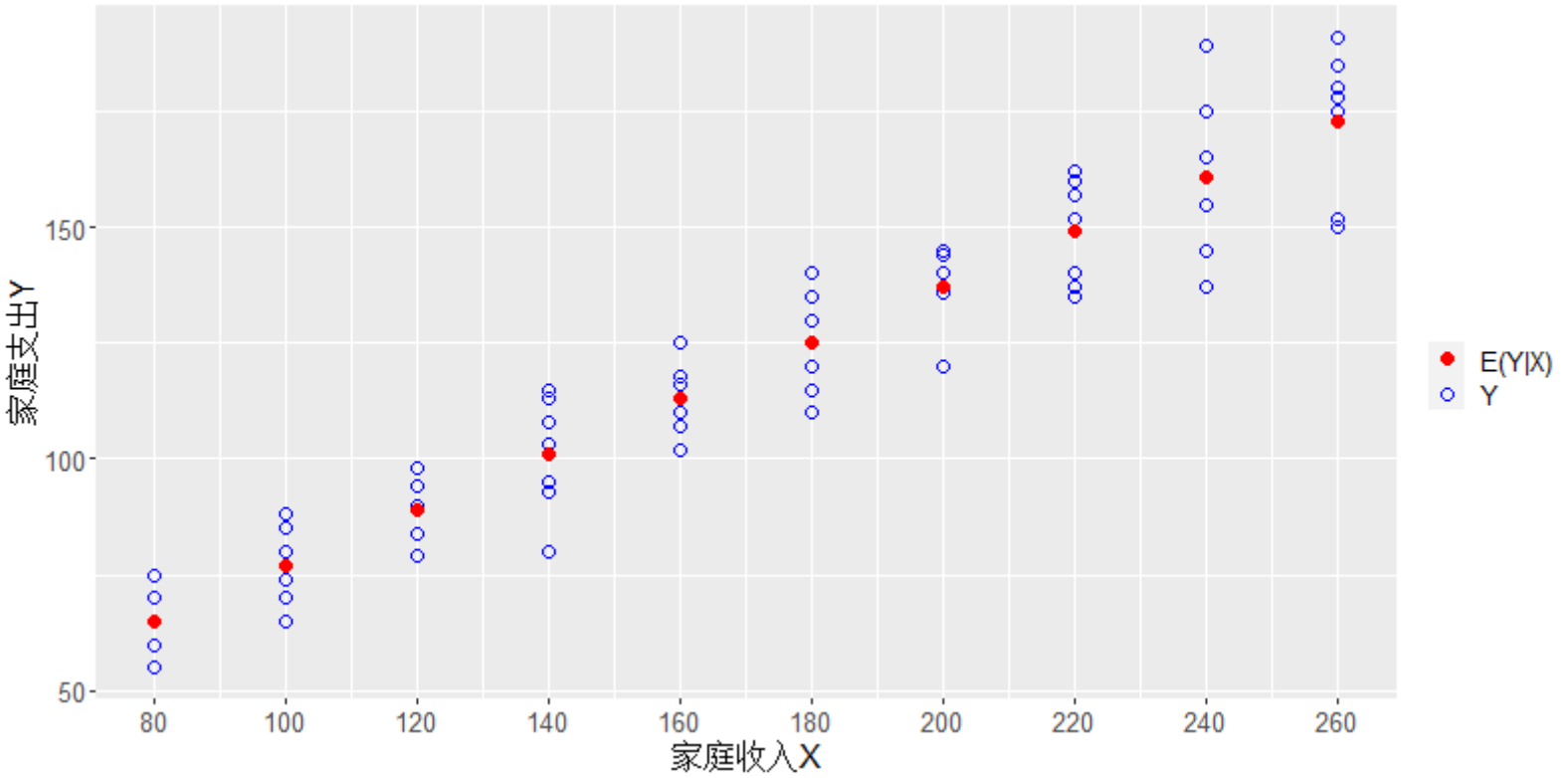
条件期望的计算过程

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

假想总体的全部数据展示

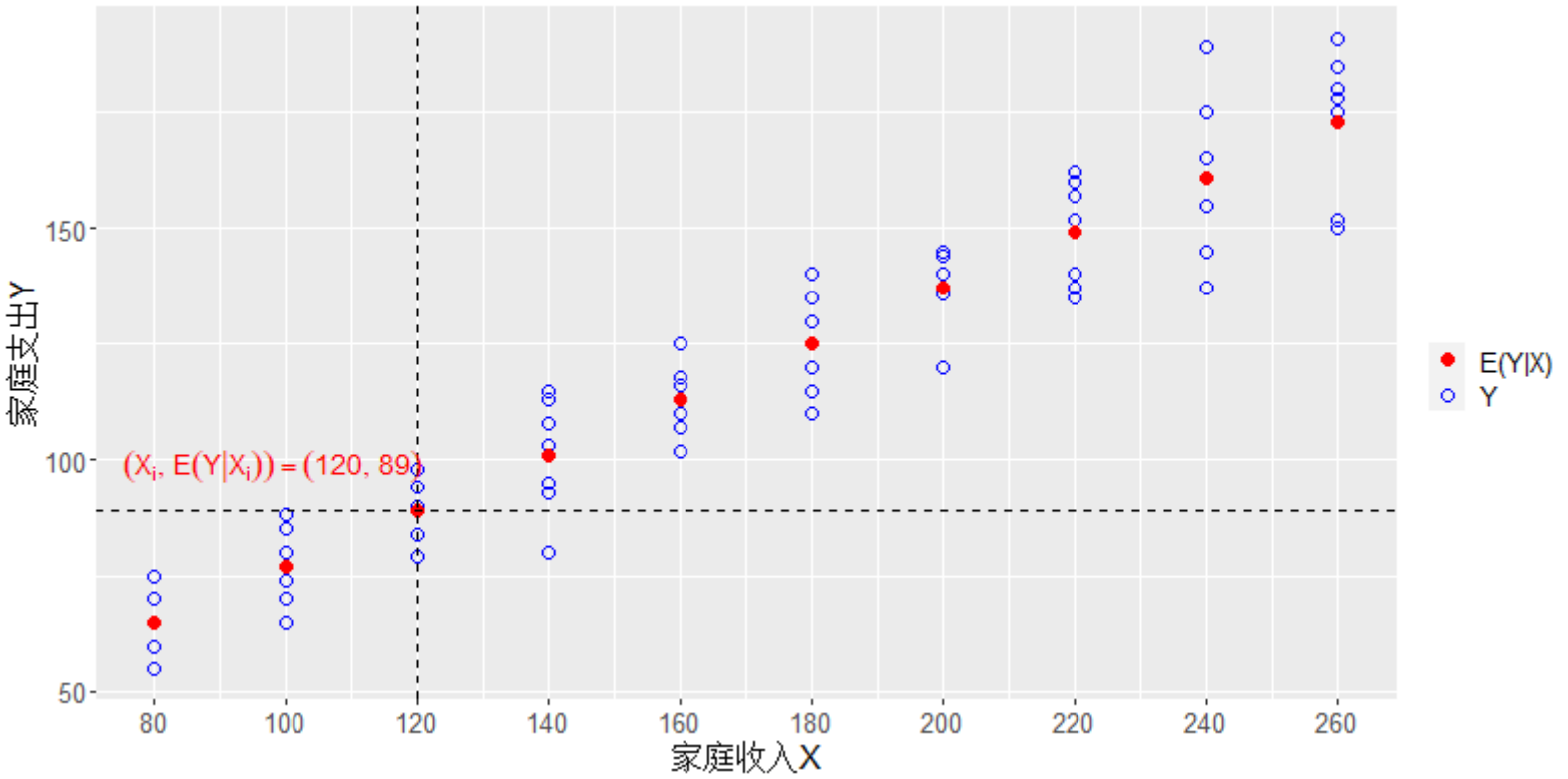


给定不同X水平下Y条件期望值



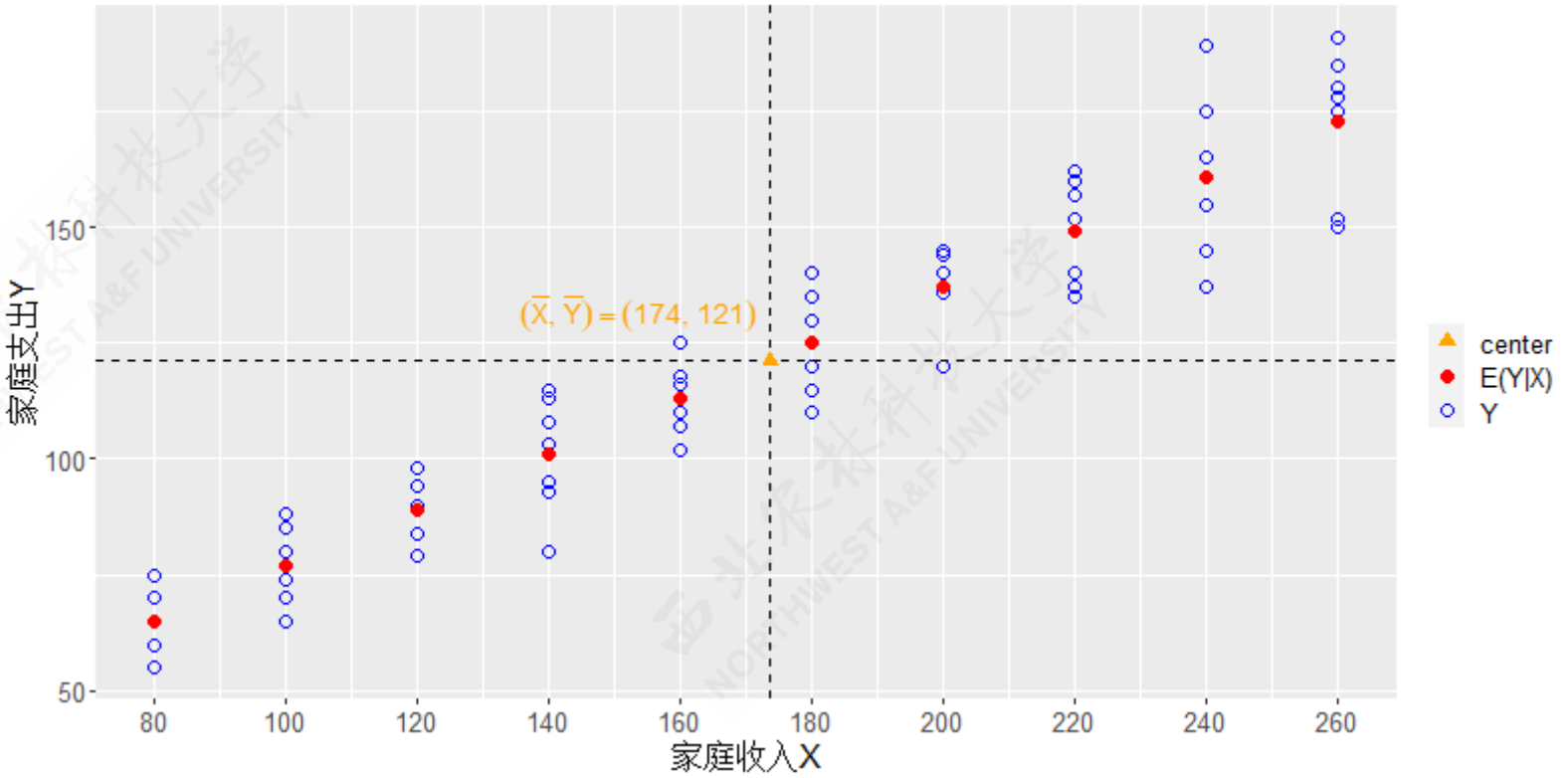
var	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
X	80	100	120	140	160	180	200	220	240	260
E(Y|X)	65	77	89	101	113	125	137	149	161	173

给定不同X水平下Y条件期望值



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

X均值和Y的无条件期望值



X均值和Y的无条件期望值

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

2.6 总体回归

总体回归线

总体回归函数

总体回归模型

随机干扰项

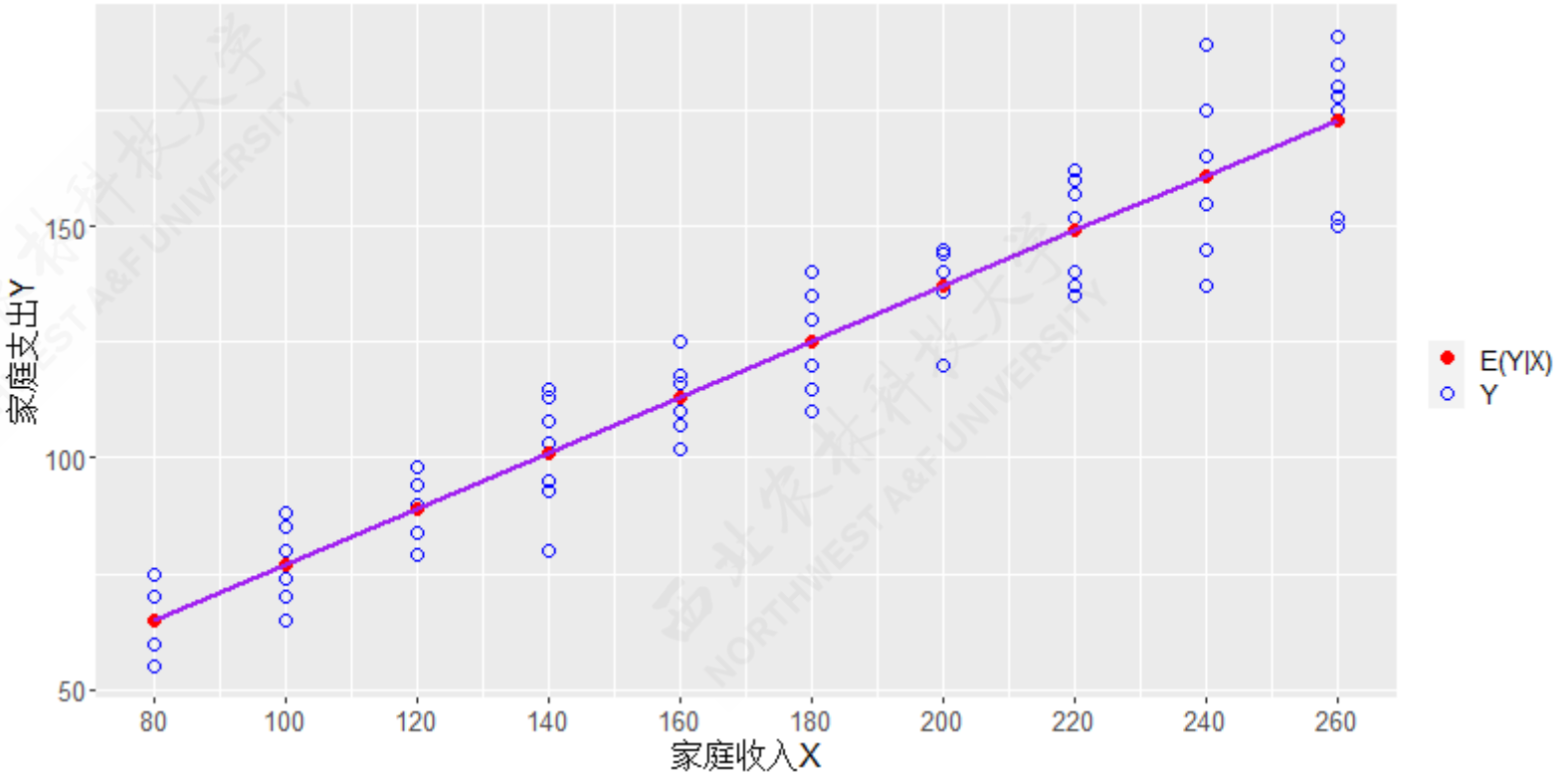
总体回归线（PRL）

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

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

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

总体回归线（PRL）



总体回归线PRL

总体回归函数（PRF）

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

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

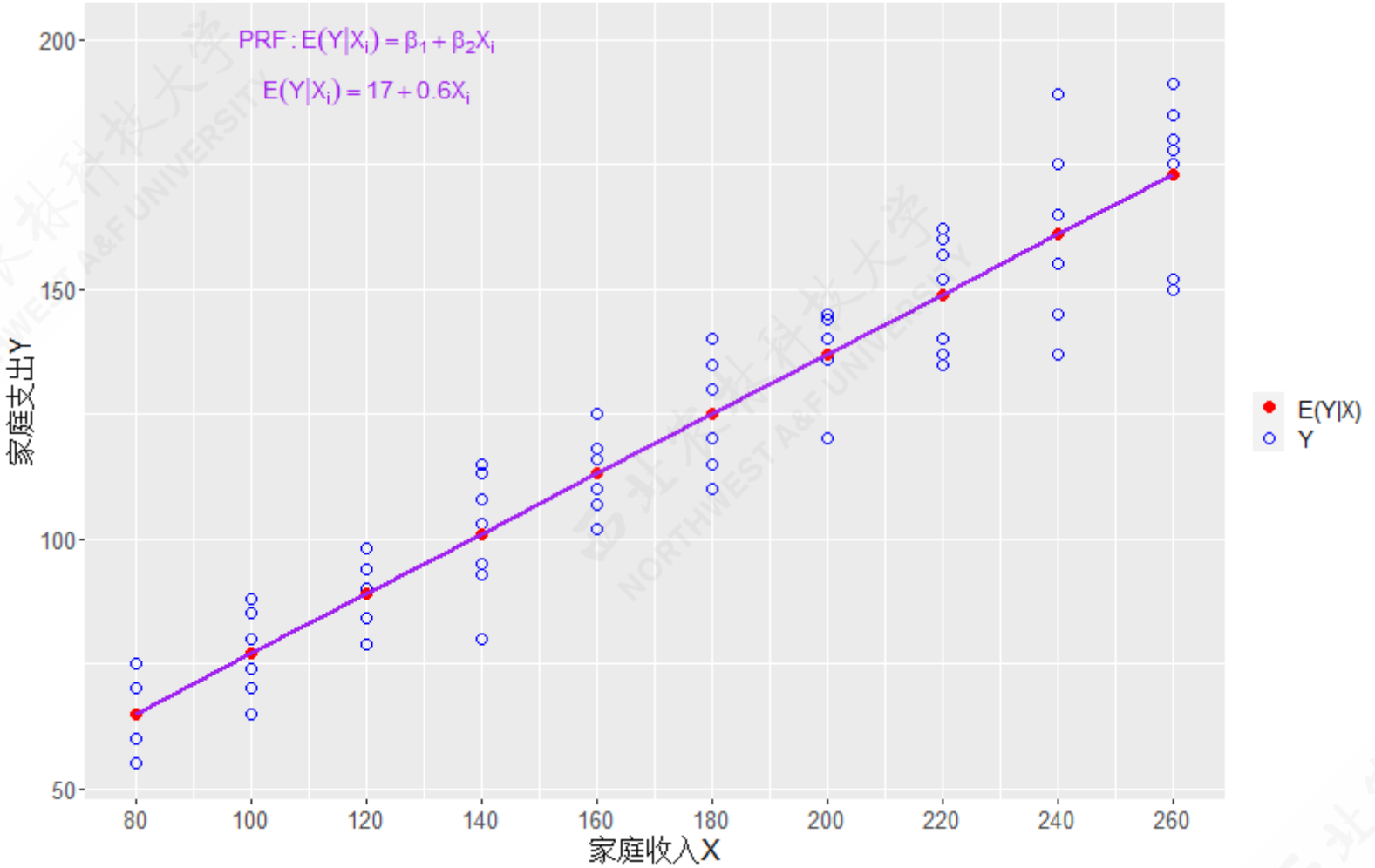
$$E(Y|X_i) = f(X_i) \qquad \text{(PRF)}$$

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

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

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

总体回归函数（PRF）



总体回归线PRL与总体回归函数PRF

总体回归模型（PRM）

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

如果总体回归函数为隐函数，则**总体回归模型**记为：

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

如果总体回归函数为线性函数，则**总体回归模型**记为：

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

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

随机干扰项

总体回归模型（PRM）设定下，Y_i将由两个部分组成。

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

随机干扰项：

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

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

随机干扰项

随机干扰项的来源：

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

随机干扰项

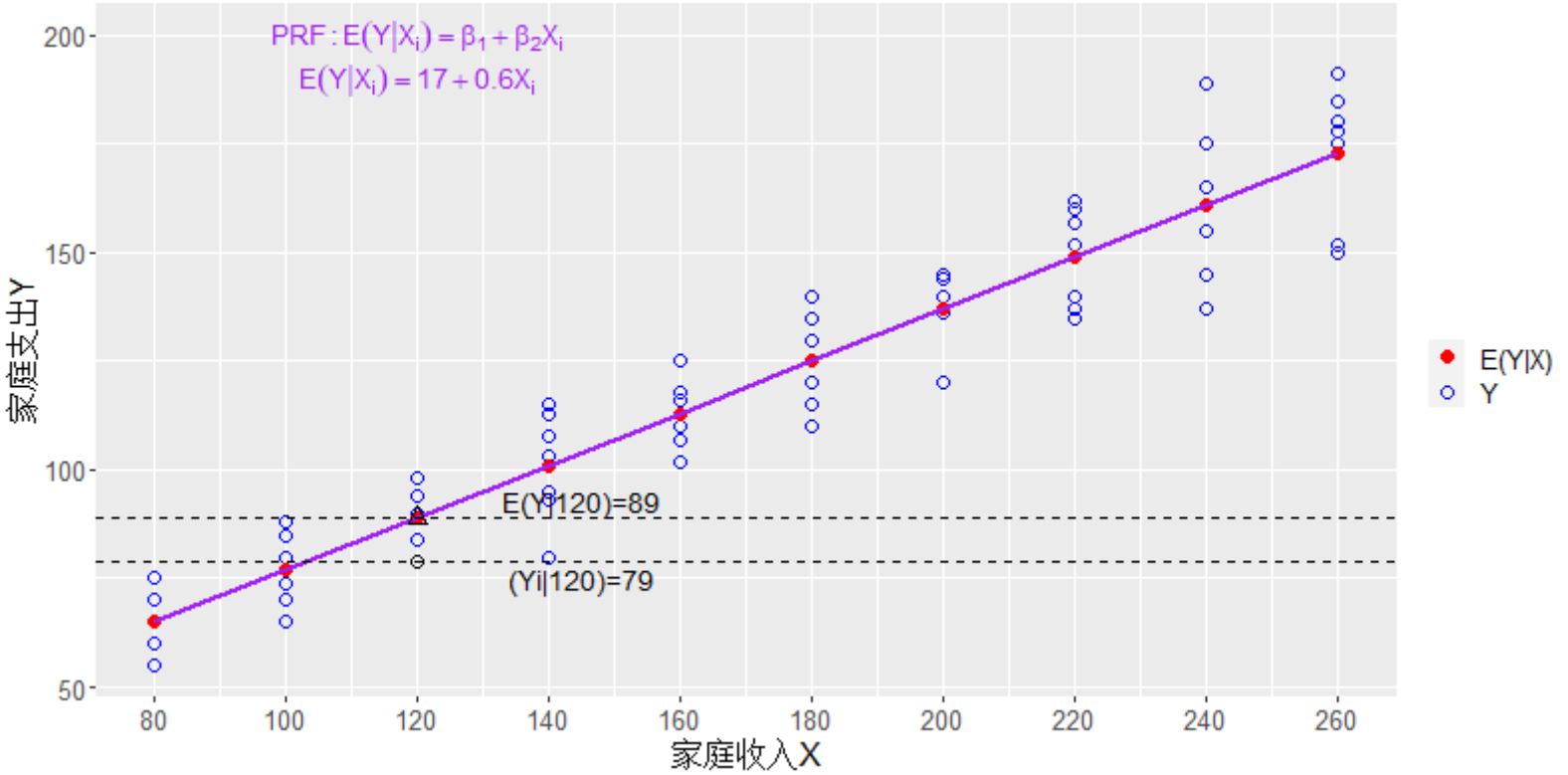
为何是“随机的”?

- 测不准？（误差）
- 测错了？（误导）
- 免不了!（内在性）

拥抱随机世界

- 风筝： Y_i
- 风筝线： $E(Y|X_i)$
- 风： u_i

理解PRM和PRF的关系



若给定一个特定家庭$(X_i = 120, Y_i = 79)$。

给定条件下，条件期望为$E(Y|120) = 89$

理解PRM和PRF的关系

若给定 $X_i = 120$，则5个家庭的真实消费支出分别为：

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

理解PRM和PRF的关系

主要结论：

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

课后思考：

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

“线性回归模型”中“线性”一词的含义

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

测试题

下列模型分别属于哪一类？请指出来：

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

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

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

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

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

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

测试题

下列模型分别属于哪一类？请指出来：

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

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

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

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

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

2.7 样本回归

样本回归线

样本回归函数

样本回归模型

残差

样本回归线(SRL)

样本(Sample)：

- 从总体中随机抽取得到的数据。

样本回归线(Sample Regression Line，SRL)：

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

样本回归函数(SRF)

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

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

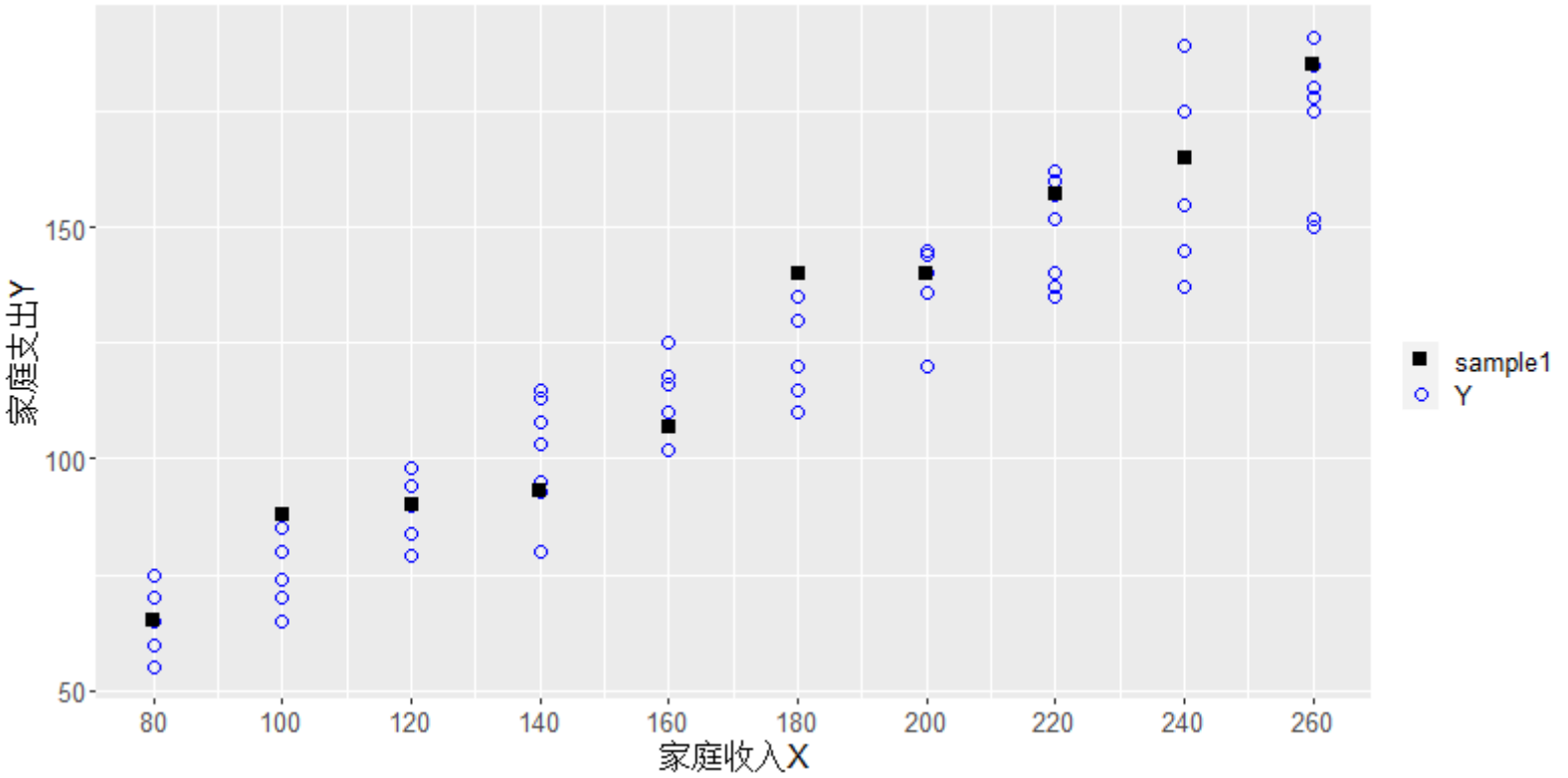
对比总体回归函数（PRF）：

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

可以认为：

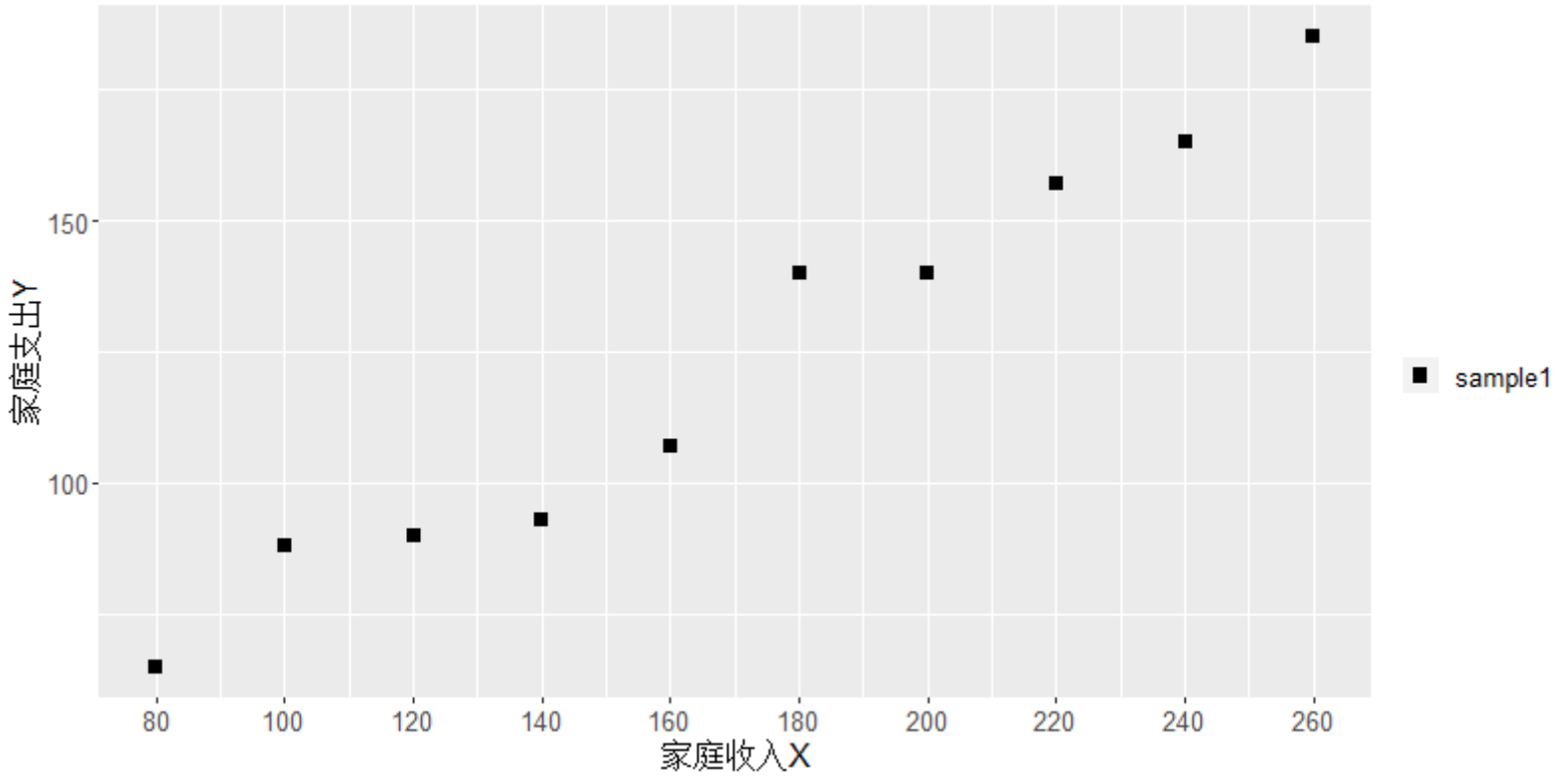
- $\hat{Y}_i$是对 $E(Y|X_i)$的估计量。
- $\hat{\beta}_1$是对 β_1的估计量。
- $\hat{\beta}_2$是对 β_2的估计量。

第一份随机样本：抽样



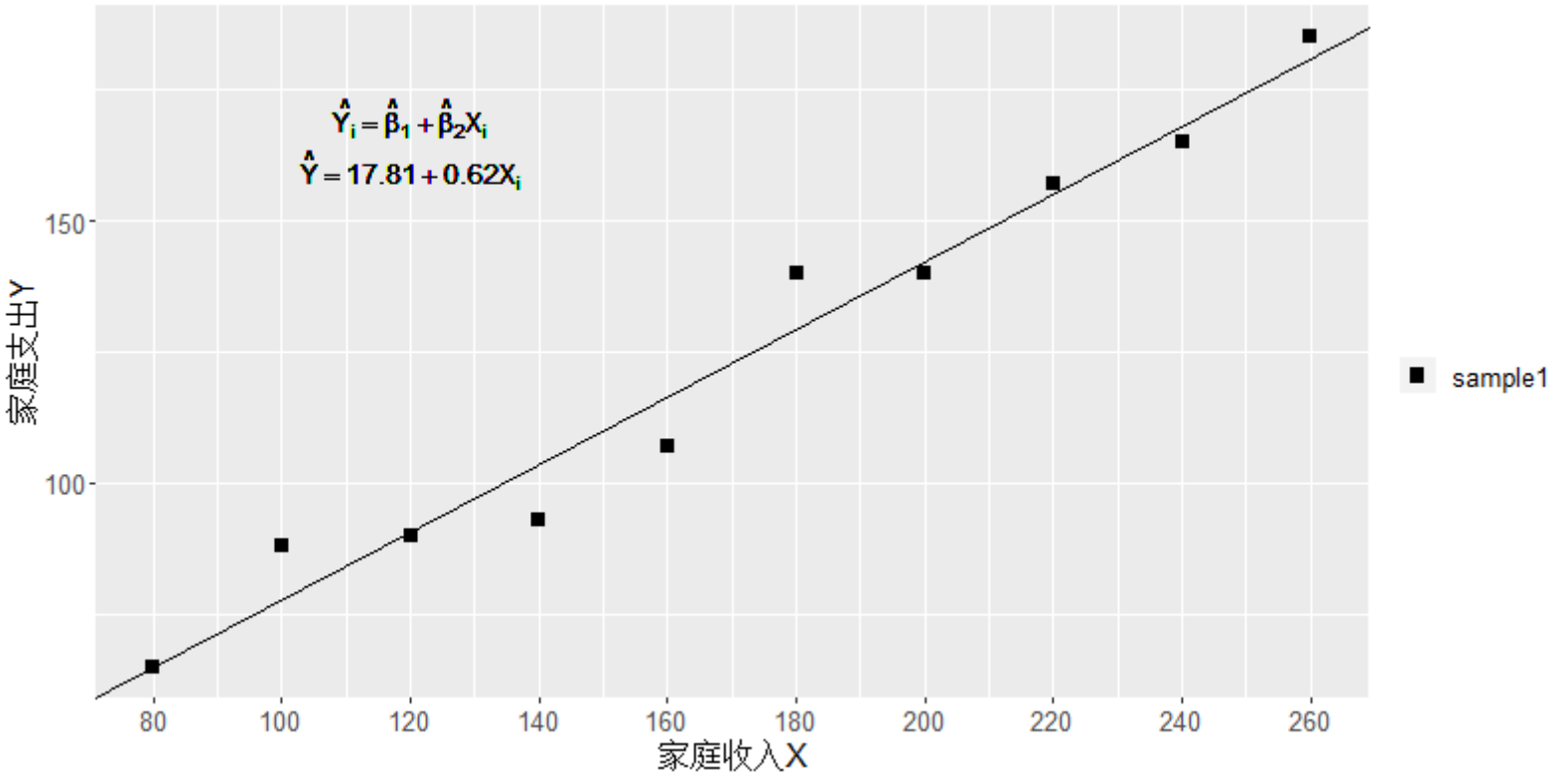
var	n1	n2	n3	n4	n5	n6	n7	n8	n9	n10
X	80	100	120	140	160	180	200	220	240	260
Y	65	88	90	93	107	140	140	157	165	185

第一份随机样本：数据



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

第一份随机样本：SRL



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

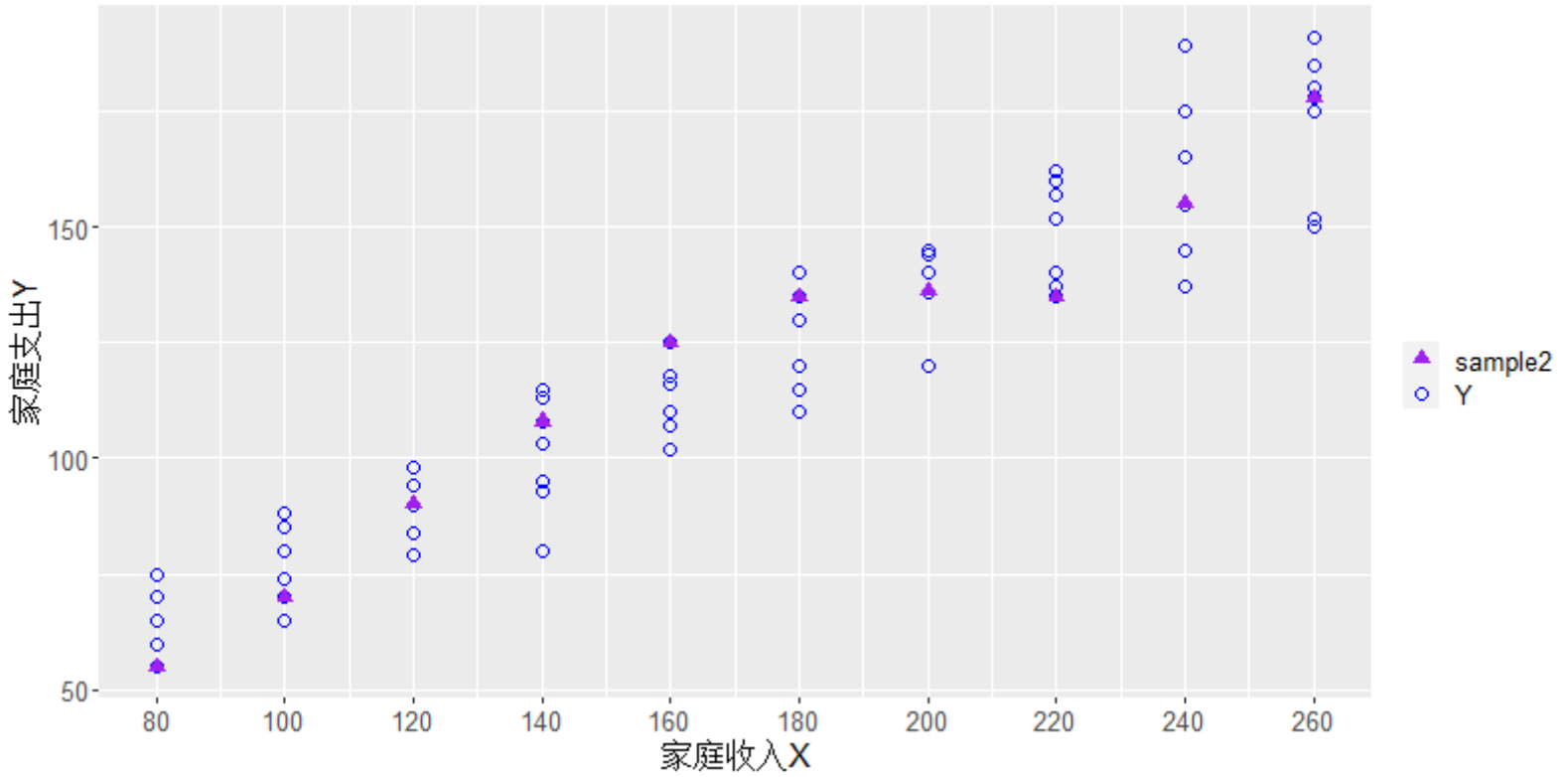
根据第一份随机样本拟合得到的**样本回归函数**SRF：

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

样本数据如下：

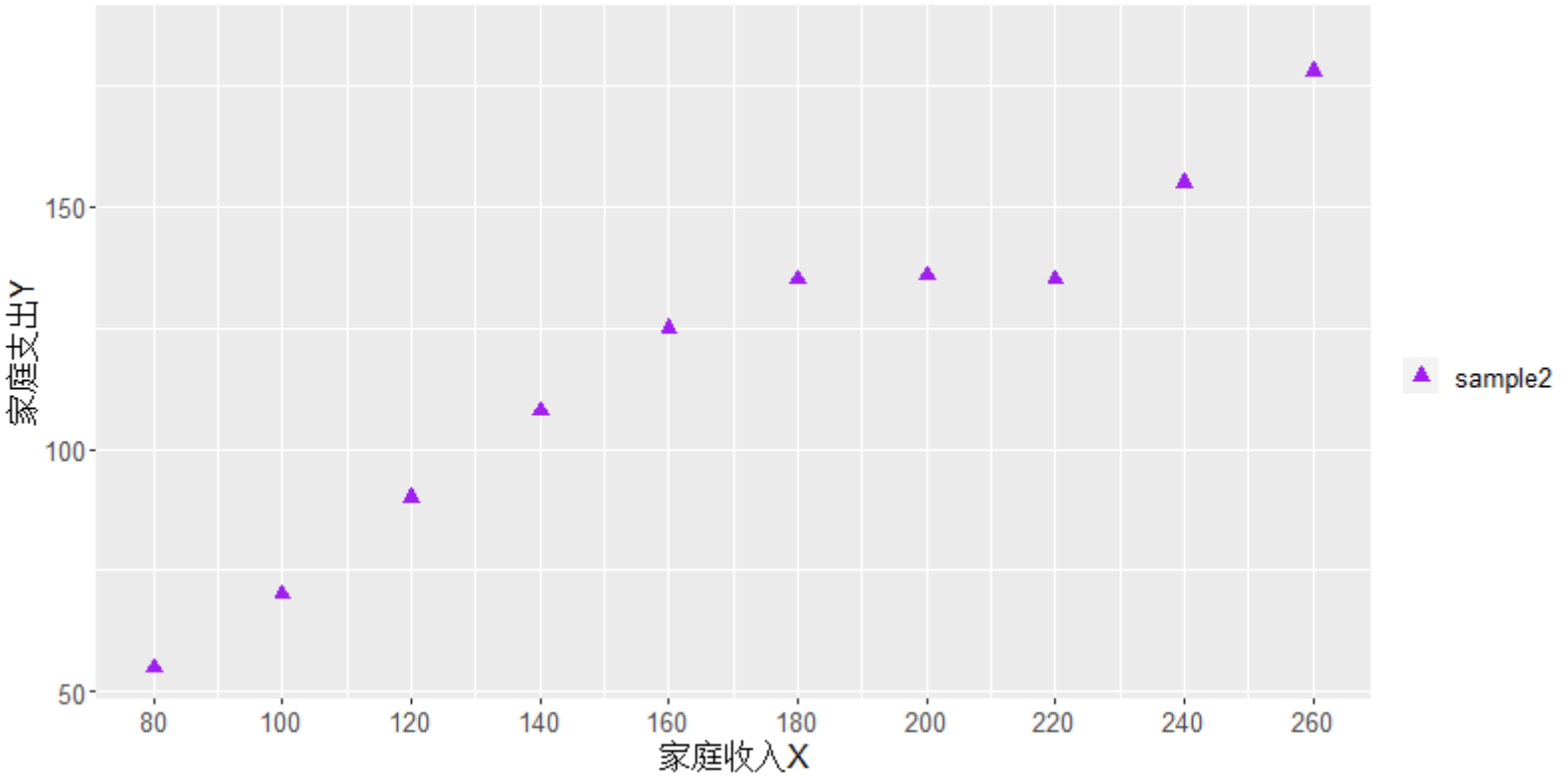
var	n1	n2	n3	n4	n5	n6	n7	n8	n9	n10
X	80	100	120	140	160	180	200	220	240	260
Y	65	88	90	93	107	140	140	157	165	185

第二份随机样本：抽样



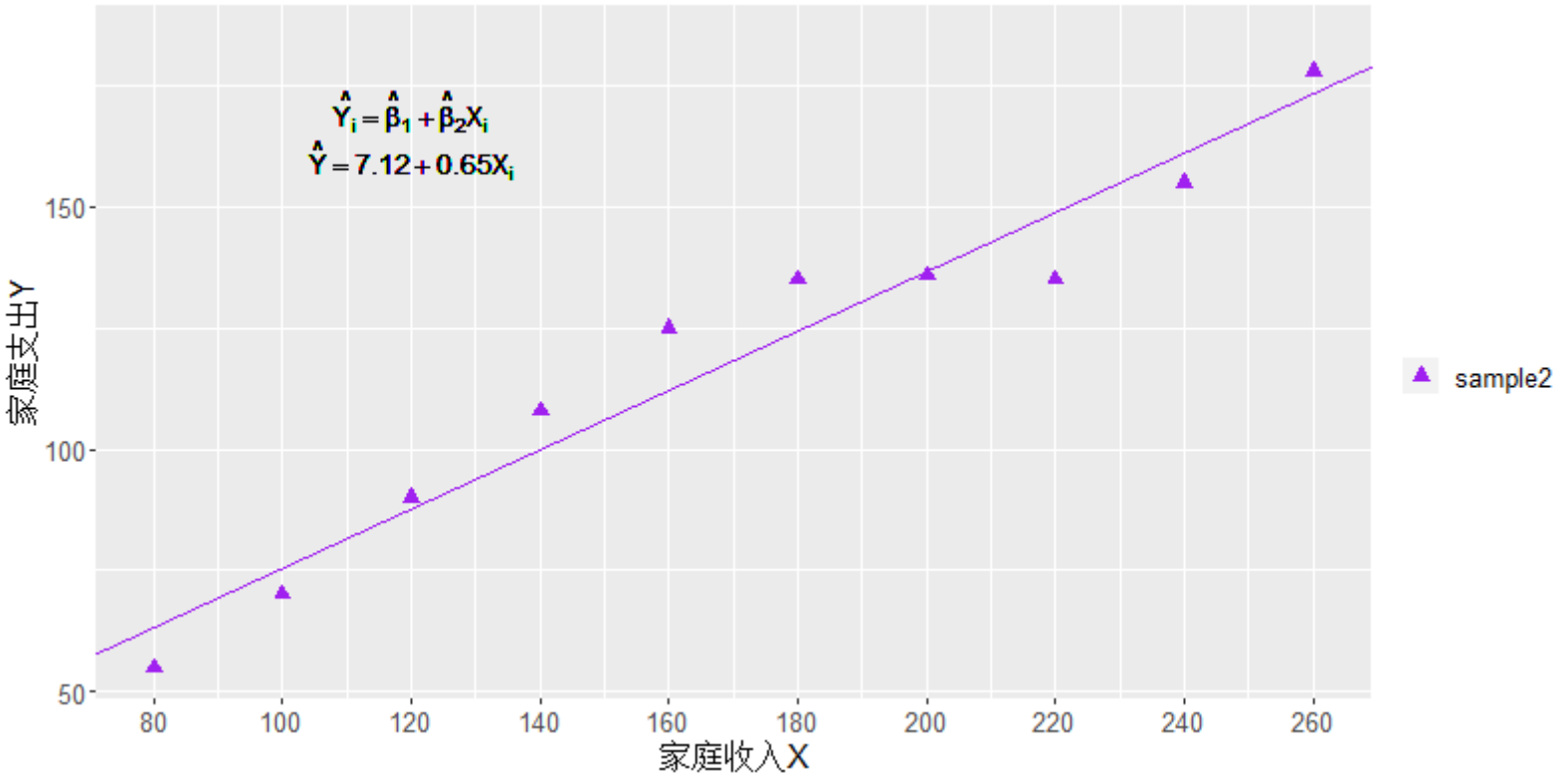
var	n1	n2	n3	n4	n5	n6	n7	n8	n9	n10
X	80	100	120	140	160	180	200	220	240	260
Y	55	70	90	108	125	135	136	135	155	178

第二份随机样本：数据



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

第二份随机样本：SRL



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

第二份随机样本：SRF

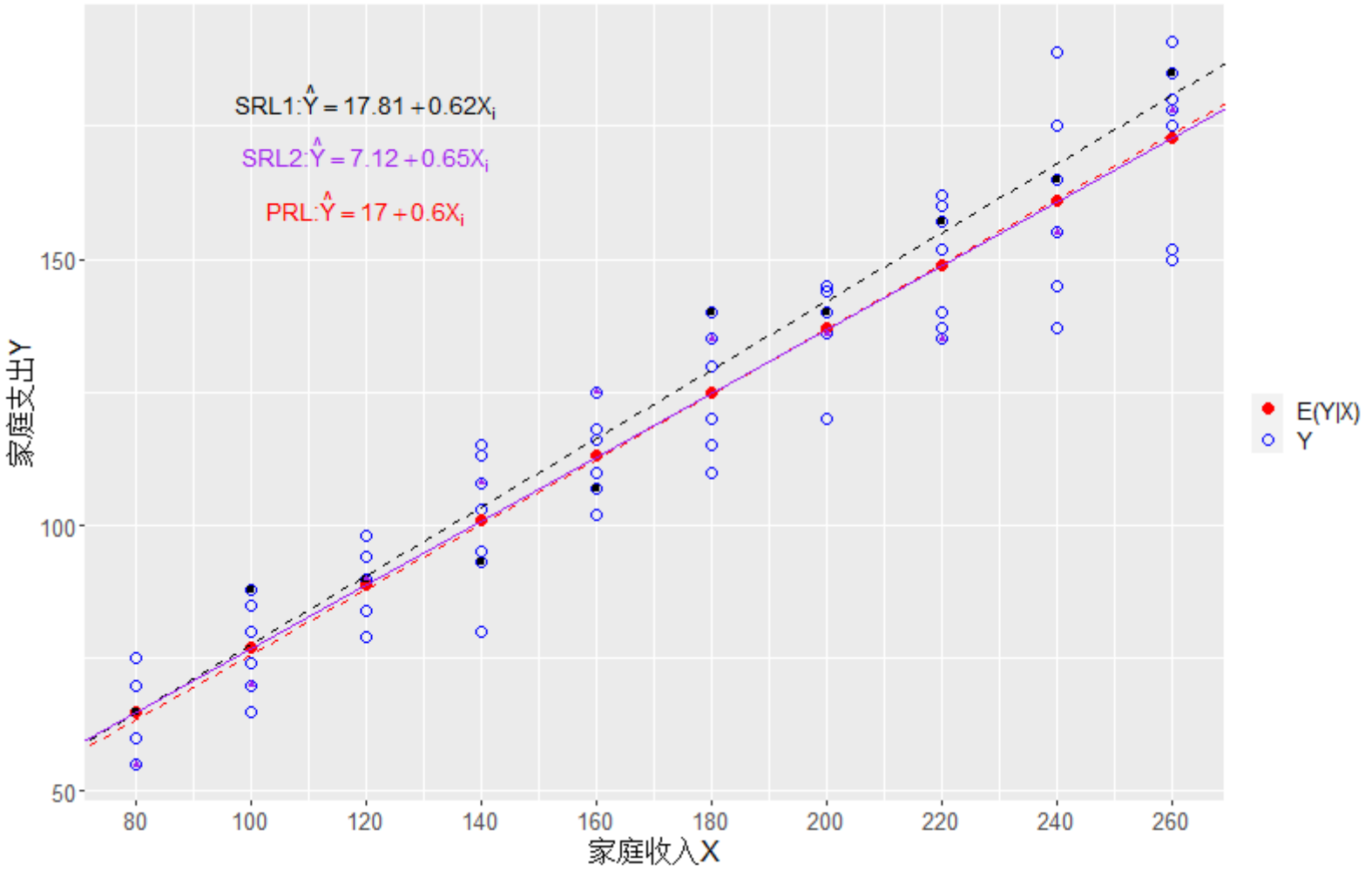
根据第二份随机样本拟合得到的**样本回归函数**SRF:

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

样本数据如下:

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

两份样本同时出现：比较分析



样本回归模型（SRM）

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

- 如果样本回归函数为隐函数，则样本回归模型可记为：

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

- 如果样本回归函数表现为直线，则样本回归模型可记为：

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

其中，e_i表示残差（Residual）

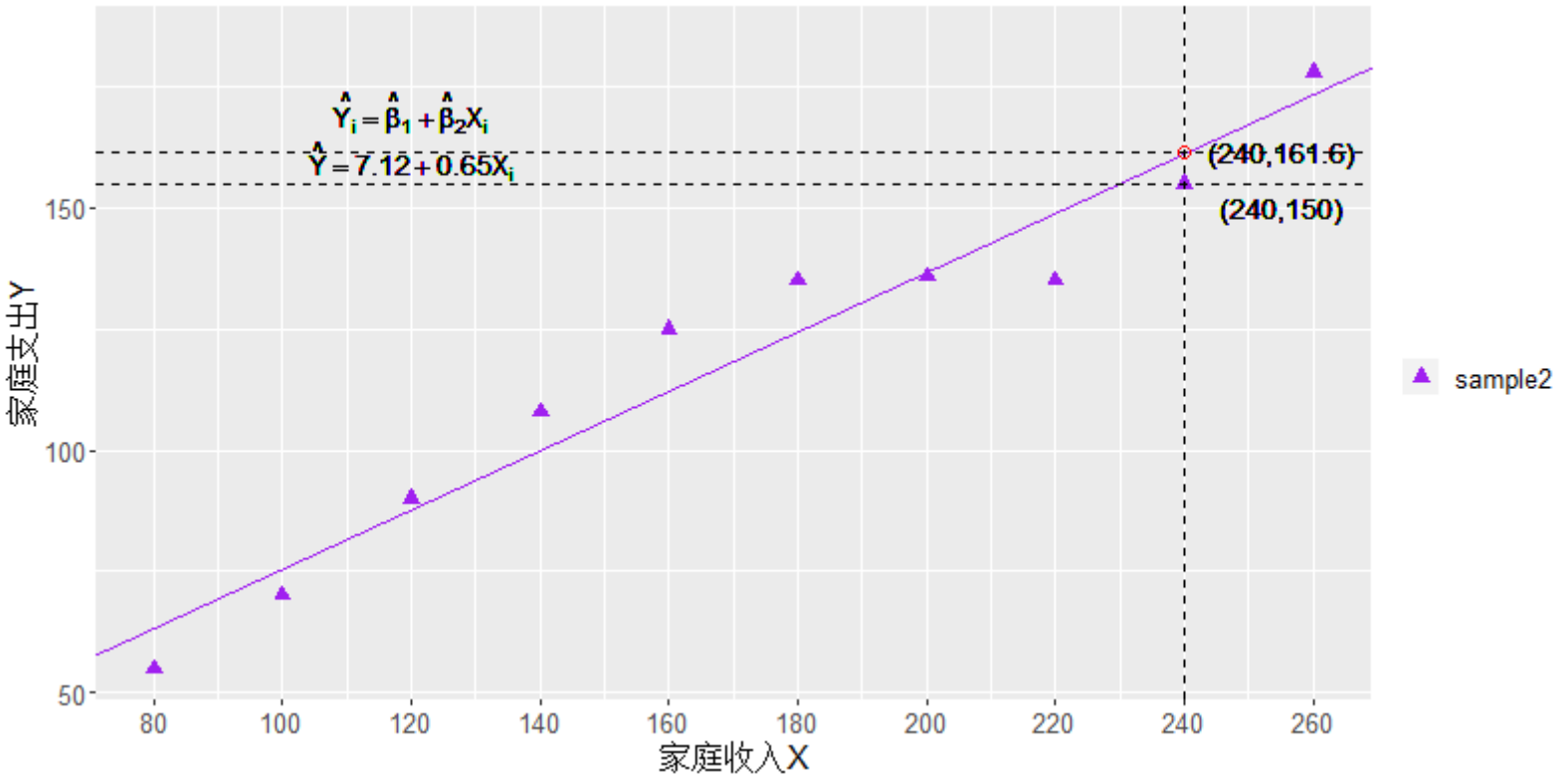
残差

残差（Residual）：

- 定义：是样本回归函数与Y的样本观测值之间的离差。
- 记号：

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

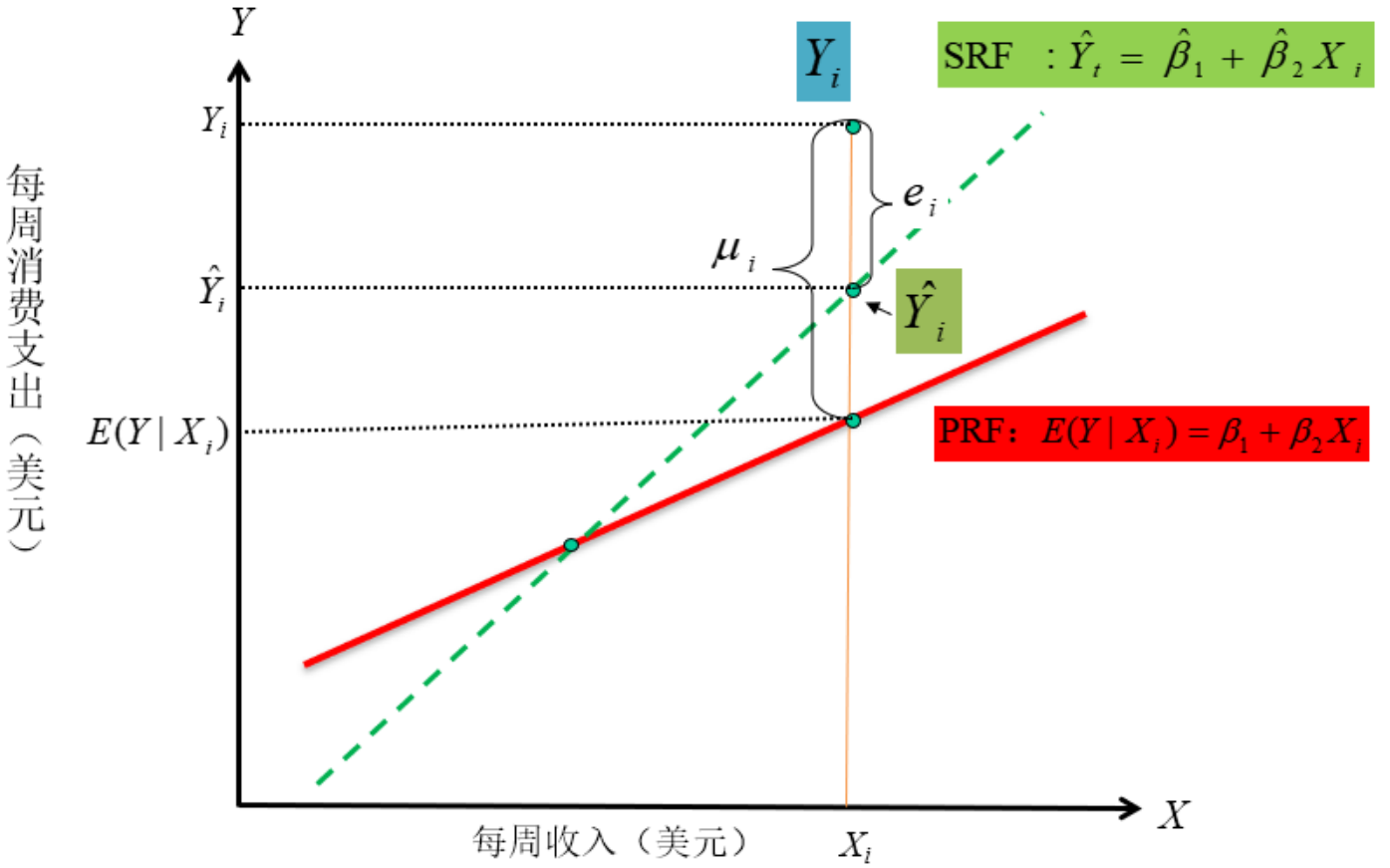
理解SRF和SRM的关系



给定 $x_i = 240$，样本2的观测值 $Y_i = 240$。 给定 $x_i = 240$，样本2的拟合值 $\hat{Y}_i = 161.6$。

残差 $e_i = Y_i - \hat{Y}_i = -6.6$。

样本回归与总体回归的比较



为何不同？继承性和变异性

样本回归与总体回归的比较

总体回归函数PRF:

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

总体回归模型PRM:

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

样本回归函数SRF:

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

样本回归模型SRM:

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

思考：

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

样本回归与总体回归的比较

知识点总结：

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

课后思考：

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

课后作业：

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

本章结束