第六章　回归分析

§6.1 相关分析回顾

§6.2 一元线性回归分析

§6.3 多重线性回归分析

相关和回归分析是研究事物的相互关系、测定它们联系的紧密程度、揭示其变化的具体形式和规律性的统计方法，是构造各种社会经济模型、进行结构分析、政策评价、预测和控制的重要工具。

§6.1　相关分析

★ 一、相关分析概述

二、相关关系的测定

比较下面两种变量间的依存关系

函数关系
（确定性关系）

1. 购物费用=单价×数量

相关关系
（非确定性关系）

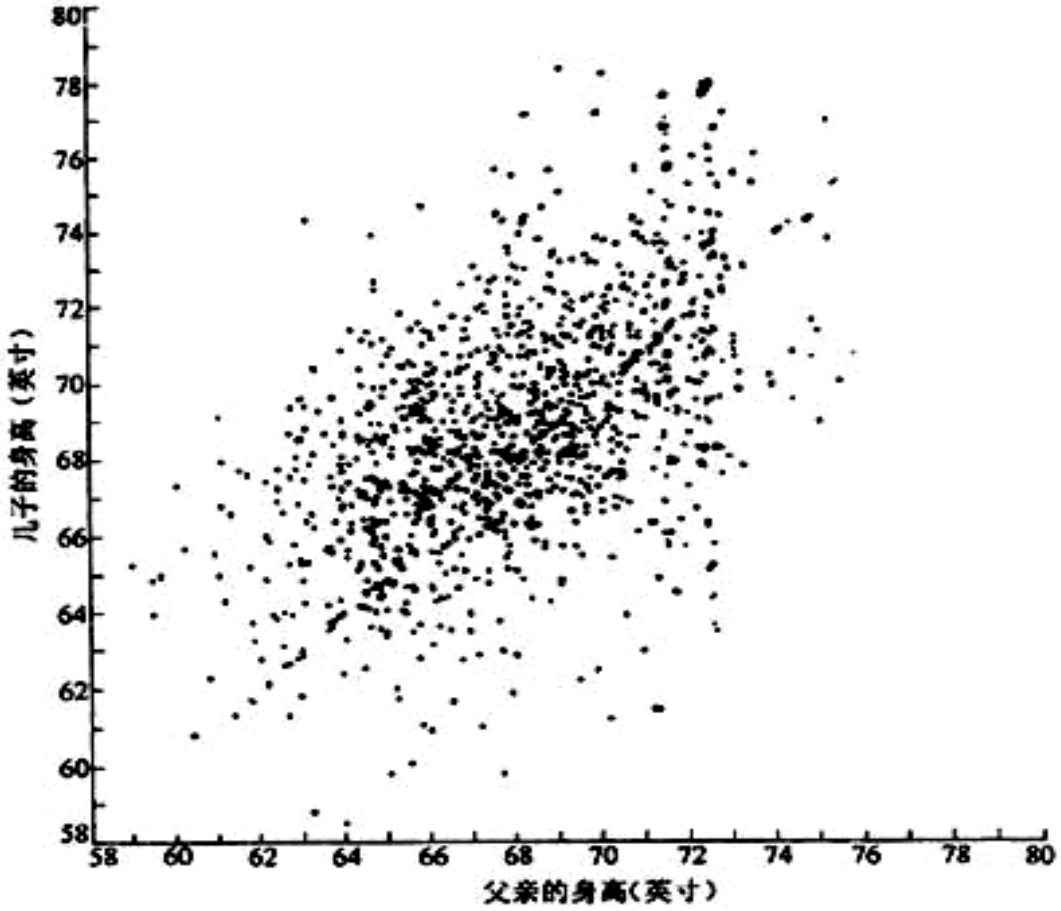
2. 个体的收入水平与其消费支出。

变量间的依存关系大致可以分成两种类型：

函数关系　指变量间所具有的严格的确定性的依存关系

相关关系　指变量间确实存在，但数量上不严格对应、无法精确表达的依存关系

函数关系与相关关系之间并无严格的界限：有函数关系的变量间，由于有测量误差及各种随机因素的干扰，可表现为相关关系；对具有相关关系的变量有深刻了解之后，相关关系有可能转化为或借助函数关系来描述。



为了研究父亲与成年儿子身高之间的关系，卡尔.皮尔逊测量了1078对父子的身高。用水平轴X上的数代表父亲身高，垂直轴Y上的数代表儿子的身高，1078个点所形成的图形是一个散点图。它的形状象一块橄榄状的云，中间的点密集，边沿的点稀少，其主要部分是一个椭圆。

相关关系的其他例子

- 商品的消费量(y)与居民收入(x)之间的关系
- 商品销售额(y)与广告费支出(x)之间的关系
- 收入水平(y)与受教育程度之间的关系(x)
- 教育发展(y)与经济发展(x)之间的关系
- 粮食产量(y)与施肥量(x_1)、降雨量(x_2)、温度(x_3)之间的关系
- ……

相关分析的种类

相关关系的种类

1.按涉及变量的多少分为
- 一元相关
- 多重相关

2.按照表现形式不同分为
- 线性相关
- 非线相关

3.按照变化方向不同分为
- 正相关
- 负相关

§6.1　相关分析

一、相关分析概述

★　二、相关关系的测定

相关关系的测定

定性分析

是依据研究者的理论知识和实践经验，对客观现象之间是否存在相关关系，以及何种关系作出判断

定量分析

在定性分析的基础上，通过编制*相关表*、绘制*相关图*、计算*相关系数*与*判定系数*等方法，来判断现象之间相关的方向、形态及密切程度

相关表

将现象之间的相互关系，用表格的形式来反映。

简单相关表

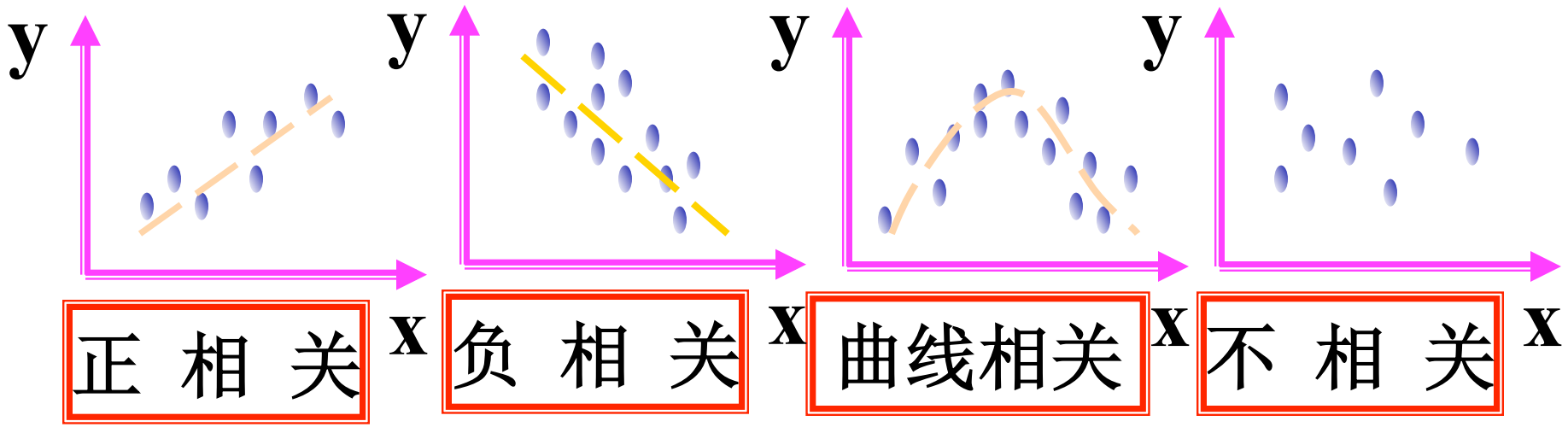
适用于所观察的样本单位数较少，不需要分组的情况

分组相关表

适用于所观察的样本单位数较多标志变异又较复杂，需要分组的情况

相关图

又称***散点图***，用直角坐标系的x轴代表自变量，y轴代表因变量，将两个变量间相对应的变量值用坐标点的形式描绘出来，用以表明相关点分布状况的图形。



相关系数

在*直线相关*的条件下，用以反映*两变量*间*线性相关*密切程度的统计指标，用*r*表示

$$r=\frac{S^{2}_{xy}}{S_{x}S_{y}}=\frac{\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)/n}{\sqrt{\sum\left(x-\bar{x}\right)^{2}/n}\cdot\sqrt{\sum\left(y-\bar{y}\right)^{2}/n}}$$

相关系数r的取值范围：-1≤r≤1

r>0 为正相关，r < 0 为负相关；

|r|=0 表示不存在线性关系；

|r|＝1 表示完全线性相关；

0<|r|<1表示存在不同程度线性相关：

- |r| < 0.4 为低度线性相关；
- 0.4≤ |r| <0.7为中度线性相关；
- 0.7≤|r| <1.0为高度线性相关。

判定系数

是相关系数的平方，用 r^2 表示；用来衡量回归方程对y的解释程度。

判定系数取值范围： $0 \le r^2 \le 1$

r^2 越接近于1，表明x与y之间的相关性越强； r^2 越接近于0，表明两个变量之间几乎没有直线相关关系.

【例】计算工业总产值与能源消耗量之间的相关系数及判定系数　资料

结论：工业总产值与能源消耗量之间存在高度的正相关关系，能源消耗量x的变化能够解释工业总产值y变化的95.2％。

$$r=\frac{\sum xy-\frac{\sum x\sum y}{n}}{\sqrt{n\sum x^2-\left(\sum x\right)^2}\sqrt{n\sum y^2-(\sum y)^2}}$$

$$=\frac{16\times 37887-916\times 625}{\sqrt{16\times 55086-916^2}\sqrt{16\times 26175-625^2}}=0.9757$$

$$r^2=\left(0.9757\right)^2=0.9520$$

§6.2 一元线性回归分析

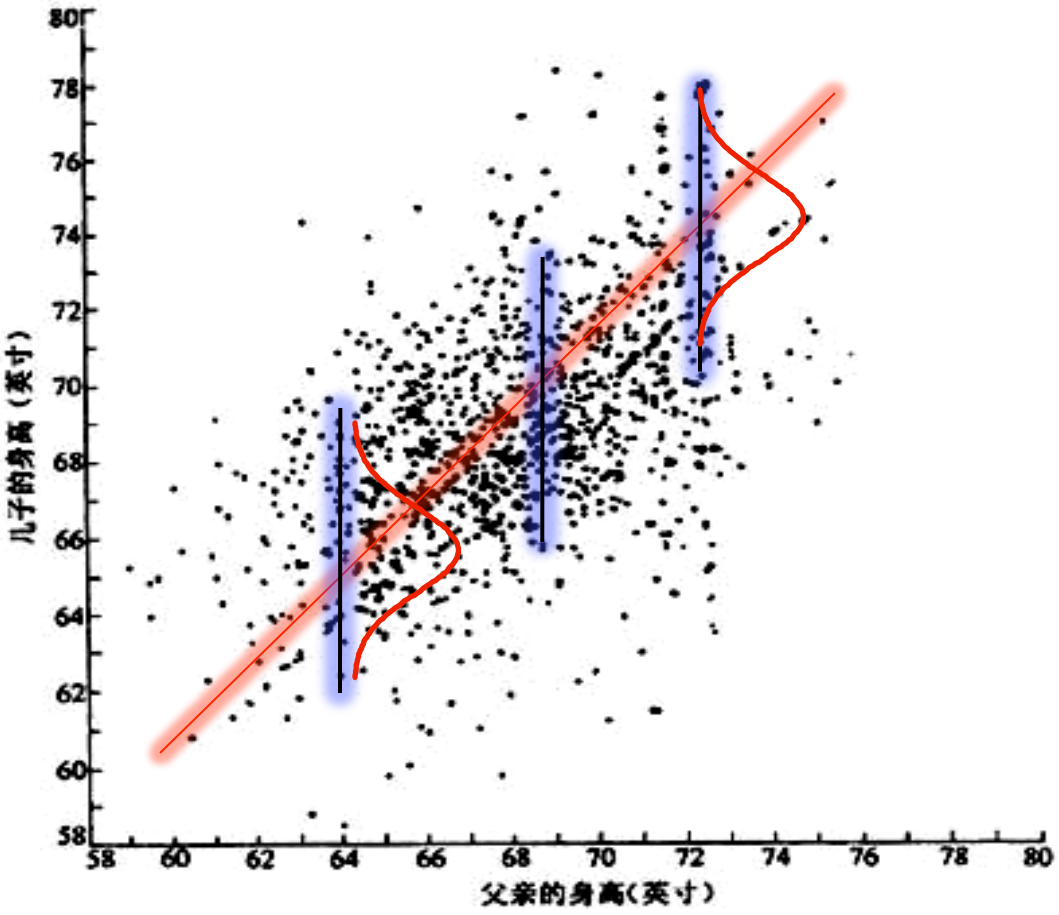
★ 一、回归分析概述

二、一元线性回归模型

三、回归估计标准差

四、线性相关的显著性检验

五、回归估计与预测


80
78
76
74
72
70
68
66
64
62
60
58
儿子的身高（英寸）
58
60
62
64
66
68
70
72
74
76
78
80
父亲的身高(英寸)

回归分析

回归：
regression

指根据相关关系的数量表达式（回归方程式）与给定的*自变量x*，揭示*因变量y*在数量上的平均变化和求得因变量预测值的统计分析方法。

回归分析的种类

Simple Linear regression

1. 按自变量的个数分
   - 一元回归（简单回归）
   - 多重回归（复回归）
2. 按回归曲线的形态分
   - 线性回归
   - 非线性回归

一元线性回归

§6.2 一元线性回归分析

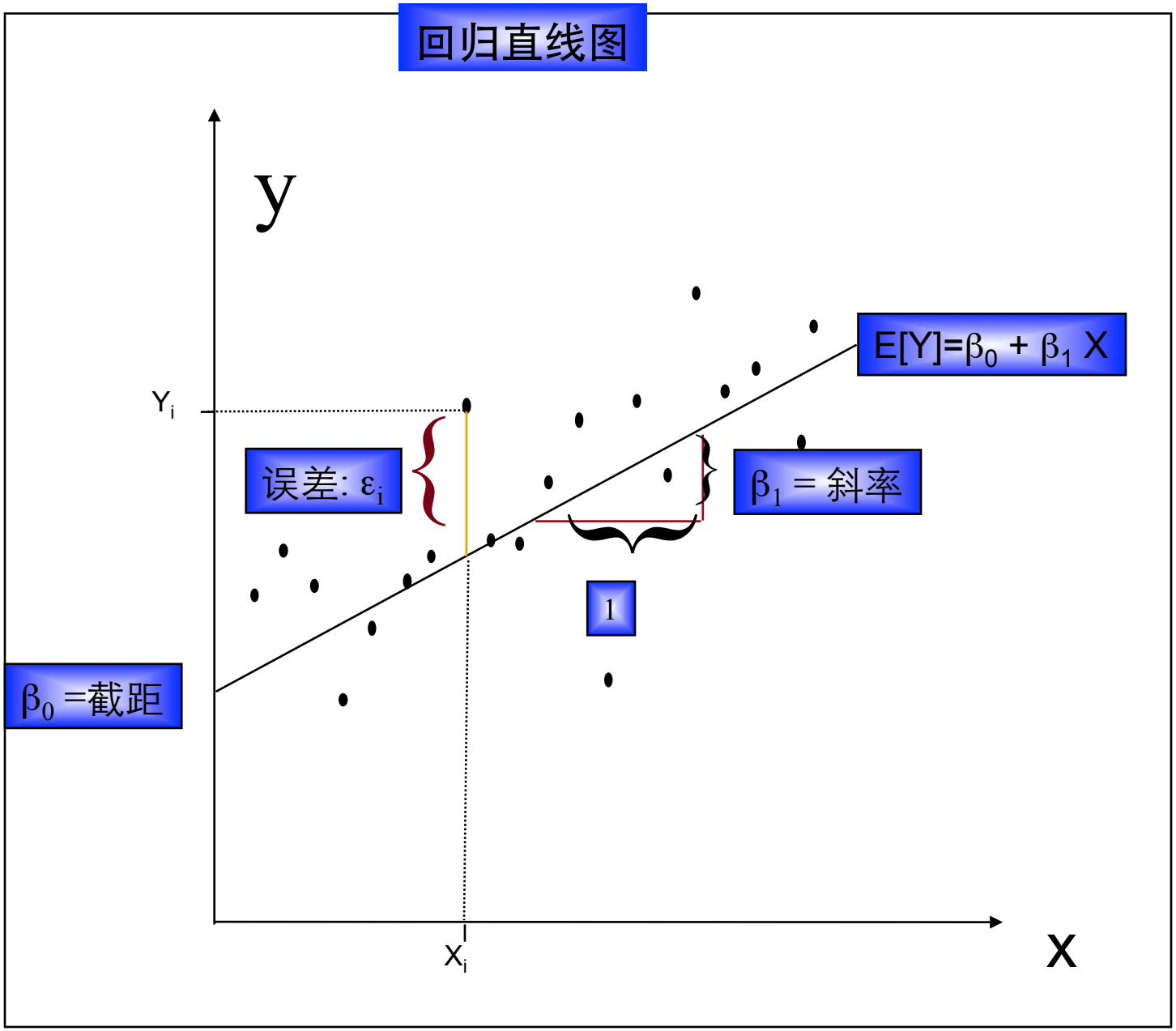
一、回归分析概述

★ 二、一元线性回归模型

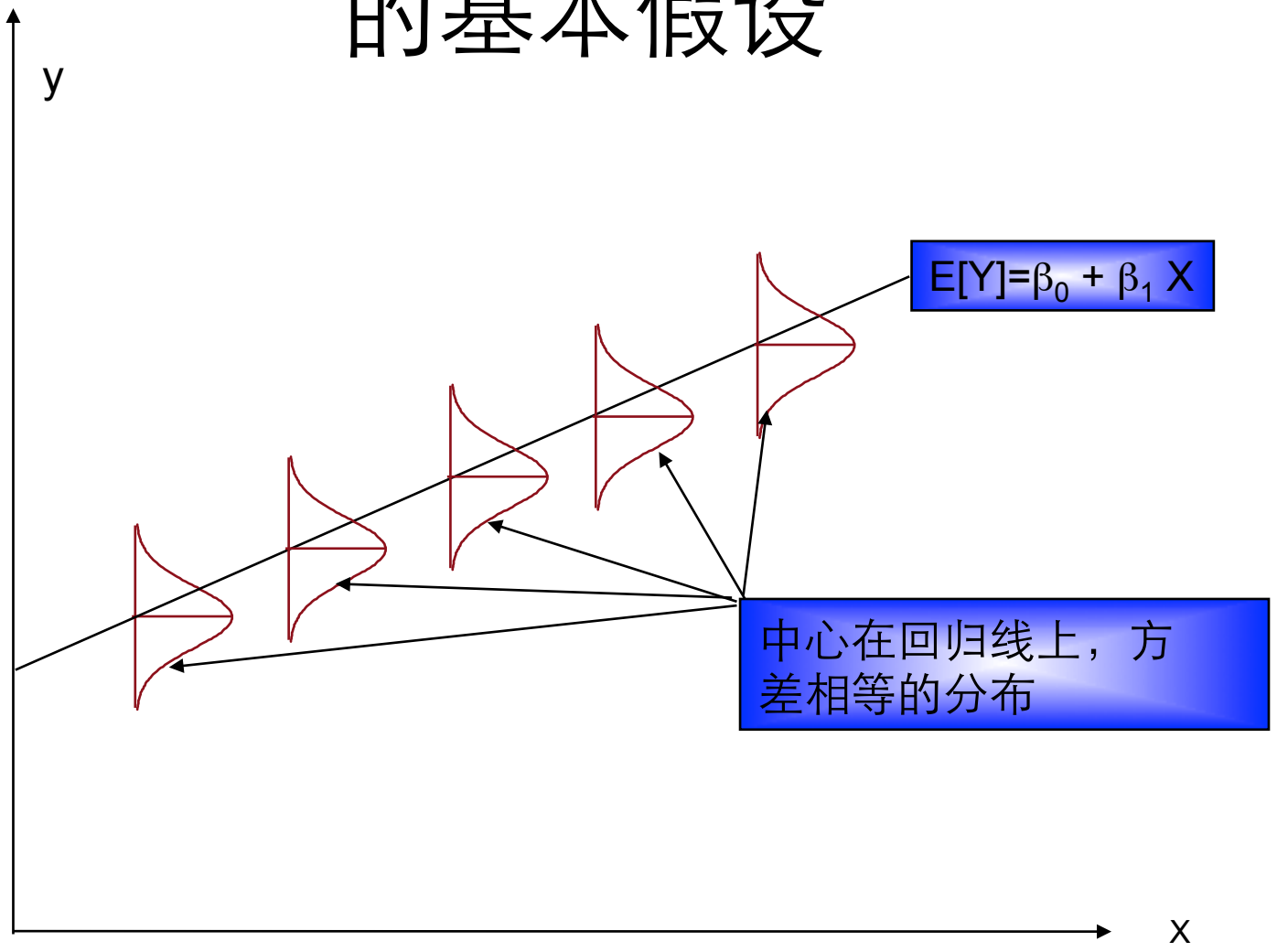
三、回归估计标准差

四、线性相关的显著性检验

五、回归估计与预测


回归直线图
y
Y_i
误差: ε_i
$E[Y]=\beta_0 + \beta_1 X$
β_1 = 斜率
1
β_0 =截距
X_i
X

简单线性回归的基本假设



一元线性回归模型

对于经判断具有线性关系的两个变量y与x，构造一元线性回归模型为：

$$Y = \alpha + \beta X + \varepsilon$$

式中：α与β为模型参数，ε为随机误差项

假定$E(\varepsilon)=0$，有总体一元线性回归方程：

$$\hat{Y} = E(Y) = \alpha + \beta X$$

Y为被解释变量(因变量Dependent variable)

x为解释变量(自变量Independent variable)

β_0,β_1是未知参数,叫回归系数(Coefficient of regression)

ε是随机误差(不可观察)

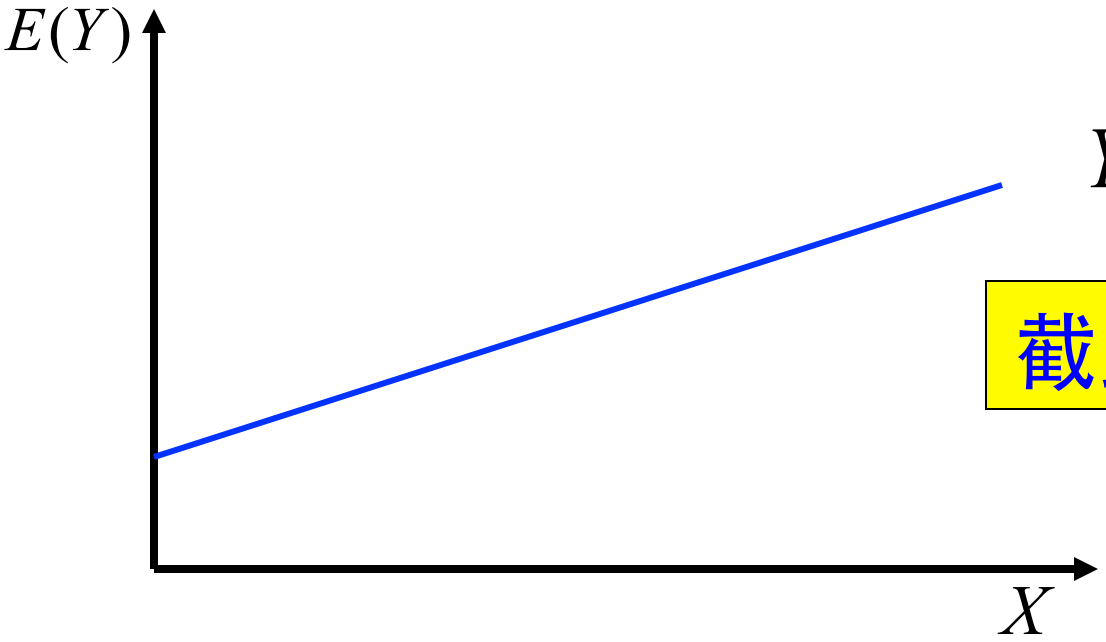
假定$E(\varepsilon)=0,\mathrm{var}(\varepsilon)=\sigma^2,\varepsilon\sim N(0,\sigma^2),$

$\Rightarrow y\sim N(\beta_0+\beta_1 x,\sigma^2),$

$E(y)=\beta_0+\beta_1 x, Var(y)=\sigma^2$

$E(y)$平均意义上表达了变量y与x统计规律性.

一元线性回归方程的几何意义

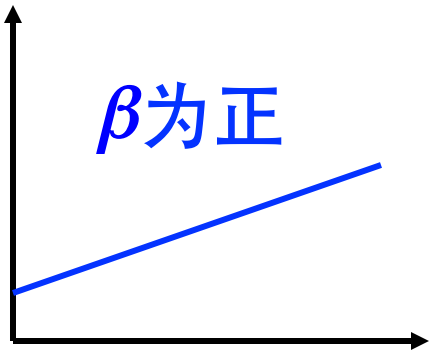


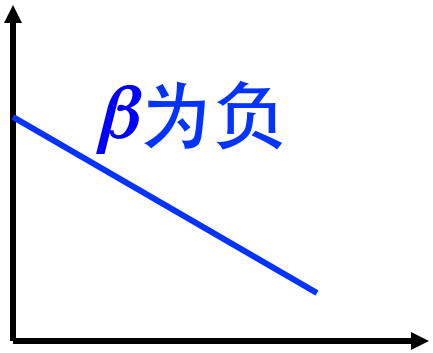
$\hat{Y} = \alpha + \beta X$

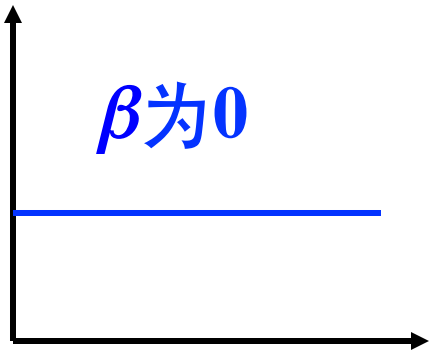
截距

斜率

一元线性回归方程的可能形态







总体一元线性回归方程：

$$\hat{Y} = E(Y) = \alpha + \beta X$$

以样本统计量估计总体参数

（估计的回归方程）

样本一元线性回归方程：

$$\hat{y} = a + bx$$

截距

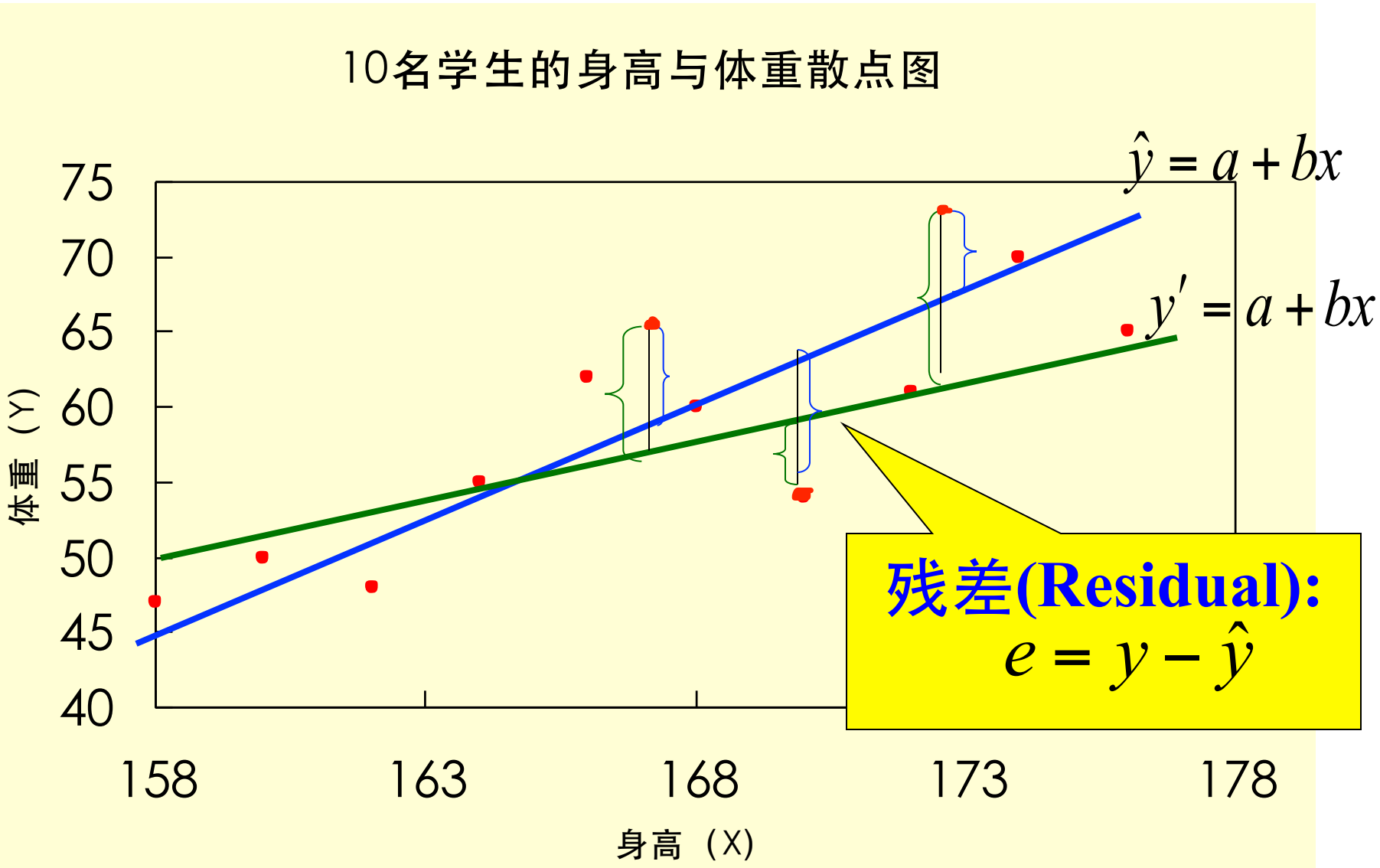
斜率（回归系数）

截距a 表示在没有自变量x的影响时，因变量y的平均水平；*回归系数b* 表明自变量x每变动一个单位，因变量y平均变动b个单位。

$\hat{y} = a + bx$是理论模型，表明x与y变量之间的平均变动关系，而变量y的实际值应为$y_i = (a + bx_i) + \varepsilon_i = \hat{y} + \varepsilon_i$

X对y的线性影响而形成的系统部分，反映两变量的平均变动关系，即本质特征。

随机干扰：各种偶然因素、观察误差和其他被忽视因素的影响


10名学生的身高与体重散点图
$\hat{y} = a + bx$
$y' = a + bx$
75
70
65
60
55
50
45
40
体重（Y）
158
163
168
173
178
身高（X）
残差(Residual):
$e = y - \hat{y}$

回归方程的求取

- 求回归方程的方法，通常是用最小二乘法，其基本思想就是从并不完全成一条直线的各点中用数理统计的方法找出一条直线，使各数据点到该直线的距离的总和相对其他任何线来说最小，即各点到回归线的差分和为最小，简称最小二乘法。

一元线性回归方程 $\hat{y} = a + bx$ 中参数a、b的确定：

最小二乘法

基本数学要求：$\sum (y - \hat{y})^2 = \min$

由$\sum (y - \hat{y})^2 = \min$，有$\sum (y - a - bx)^2 = \min,$

分别对函数中a、b求偏导数，并令其为零，有

$$\begin{cases} 2\sum (y - a - bx)(-1) = 0 \\ 2\sum (y - a - bx)(-x) = 0 \end{cases}$$

整理得到由两个关于a、b的二元一次方程组成的方程组：

$$\begin{cases} \Sigma y = na + b\Sigma x \\ \Sigma xy = a\Sigma x + b\Sigma x^2 \end{cases}$$

进一步整理，有：

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$a = \bar{y} - b\bar{x}。$$

【例】建立工业总产值对能源消耗量的线性回归方程 资料

【分析】因为工业总产值与能源消耗量之间存在高度正相关关系（$r=0.9757, r^2=0.9520$），所以可以拟合工业总产值对能源消耗量的线性回归方程。

解：设线性回归方程为 $\hat{y}=a+bx$

由计算表知$n=16, \sum x=916, \sum y=625,$

$\sum xy=37887, \sum x^2=55086,$

$$b=\frac{n\sum xy-\sum x\sum y}{n\sum x^2-\left(\sum x\right)^2}=\frac{16\times 37887-916\times 625}{16\times 55086-916^2}=0.7961$$

$$a=\bar{y}-b\bar{x}=\frac{625}{16}-0.7961\times\frac{916}{16}=-6.5142$$

即线性回归方程为：

$$\hat{y}=-6.5142+0.7961x$$

计算结果表明，在其他条件不变时，能源消耗量每增加一个单位（十万吨），工业总产值将增加0. 7961个单位（亿元）。

回归分析与相关分析

联系：

- 理论和方法具有一致性；
- 无相关就无回归，相关程度越高，回归越好；
- 相关系数和回归系数方向一致，可以互相推算。

回归分析与相关分析

区别：

- 相关分析中x与y对等，回归分析中x与y要确定自变量和因变量；
- 相关分析中x、y均为随机变量，回归分析中只有y为随机变量；
- 相关分析测定相关程度和方向，回归分析用回归模型进行预测和控制。

Attention

我们不能把回归分析看作是在变量间建立因果关系的过程。回归分析只能表明，变量是如何或者是以怎样的程度彼此联系在一起的。有关因果关系的任何结论，必须建立在理论分析的基础之上。　思考：两变量相关，是否一定存在因果联系?

§6.2 一元线性回归分析

一、回归分析概述

二、一元线性回归模型

★ 三、回归估计标准误差

四、线性相关的显著性检验

五、回归估计与预测

回归估计标准误

是因变量各实际值与其估计值之间的平均差异程度，表明其估计值对各实际值代表性的强弱；其值越小，回归方程的代表性越强，用回归方程估计或预测的结果越准确。

$$S_e = \sqrt{\frac{\sum (y-\hat{y})^2}{n-2}} = \sqrt{\frac{\sum y^2 - a\sum y - b\sum xy}{n-2}}$$

分母之所以是（ **n-2**），而不是 **n**，是因为根据样本资料用最小平方法求参数**α**和**β**时，受两个标准方程的约束，失去了两个自由度。

在大样本条件下，可用公式计算：

$$S_e=\sqrt{\frac{\sum(y-\hat{y})^2}{n}}=\sqrt{\frac{\sum y^2-a\sum y-b\sum xy}{n}}$$

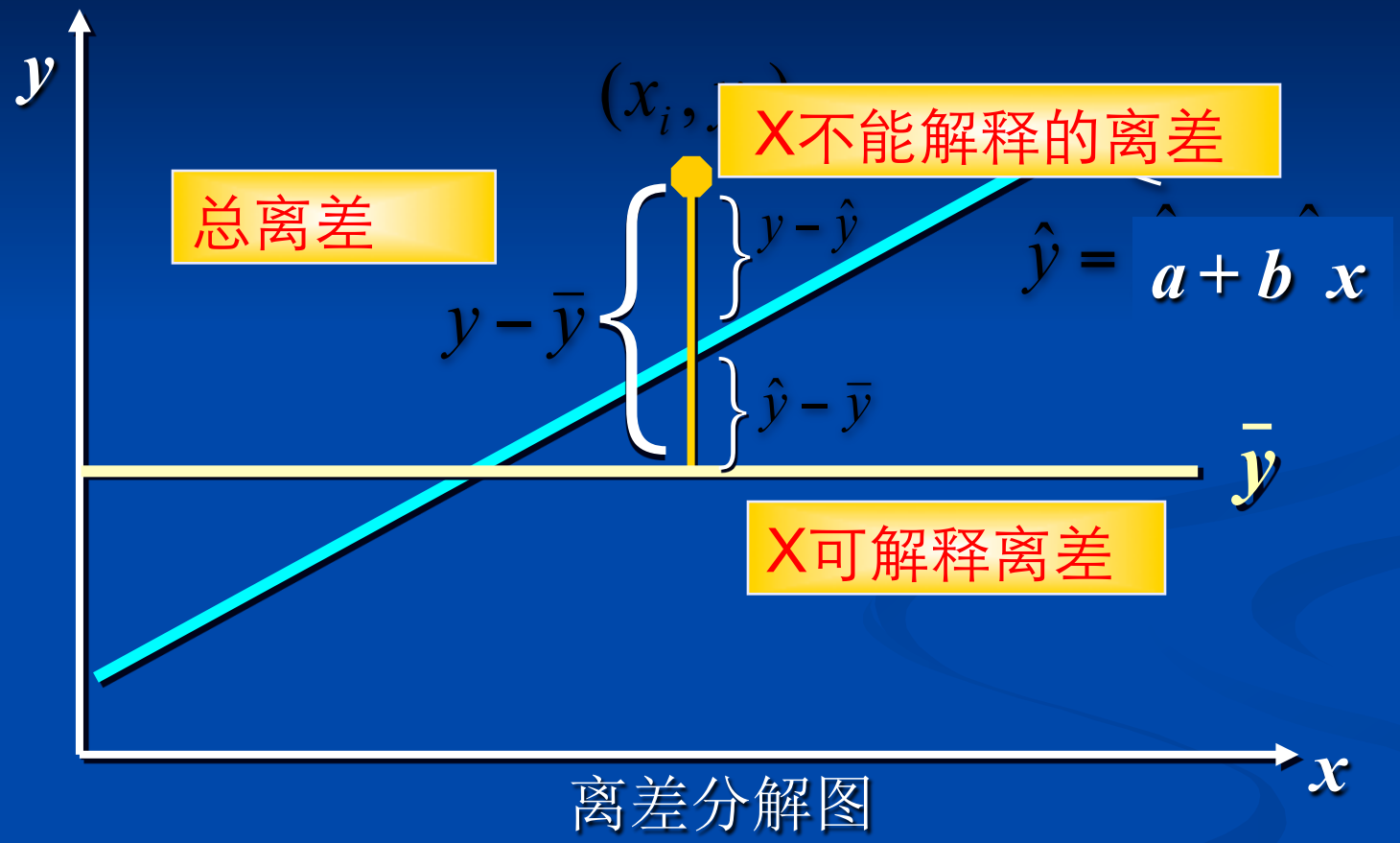
【例】计算前面拟合的工业总产值对能源消耗量回归方程的回归标准差 资料

解：已知$n=16,\sum y=625,\sum xy=37887,$

$\sum y^2=26175,$ 且知 $a=-6.5142,\quad b=0.7961$

$$S_e=\sqrt{\frac{\sum y^2-a\sum y-b\sum xy}{n-2}}=2.457(\text{亿元})$$

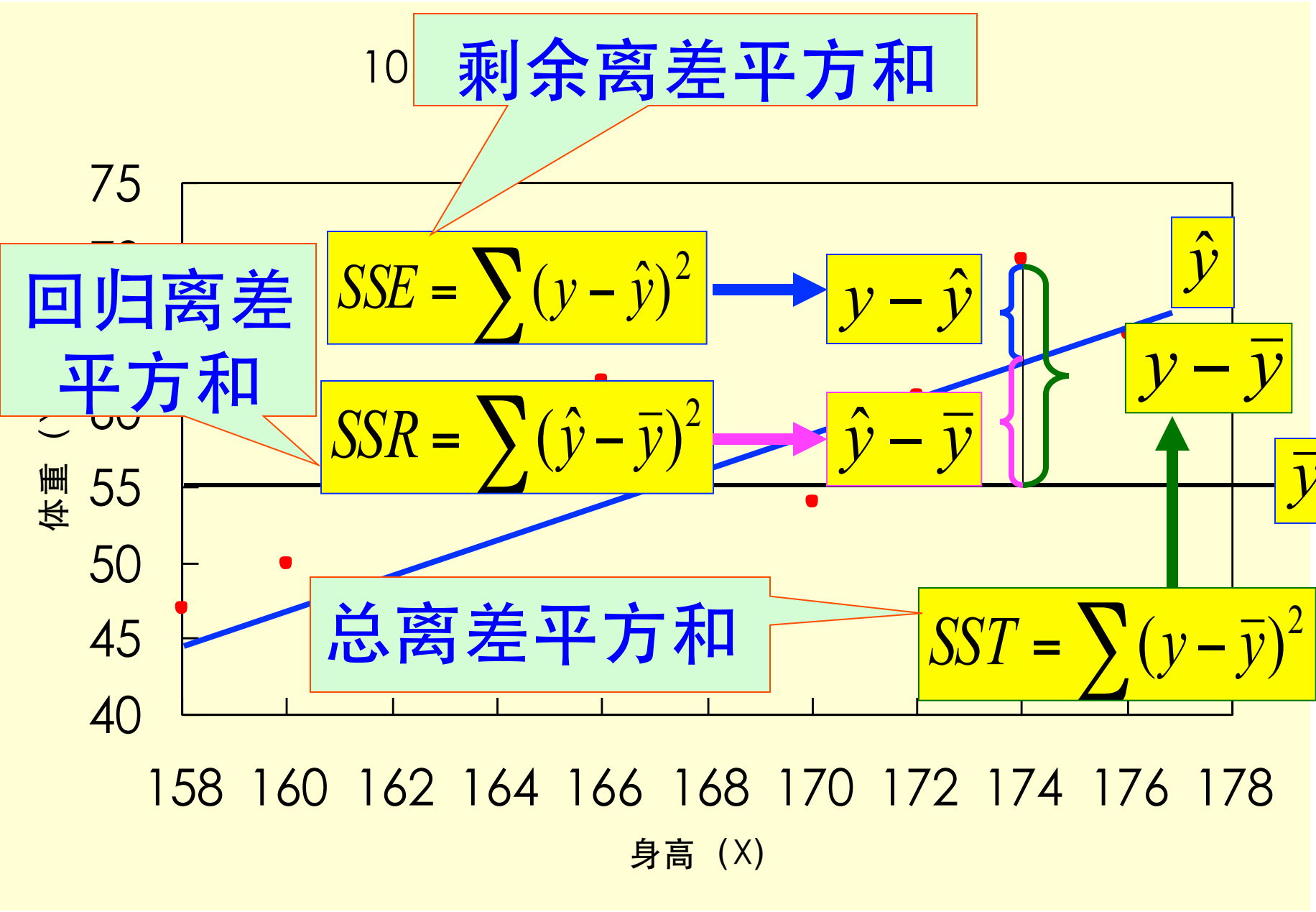
离差的分解（图示）



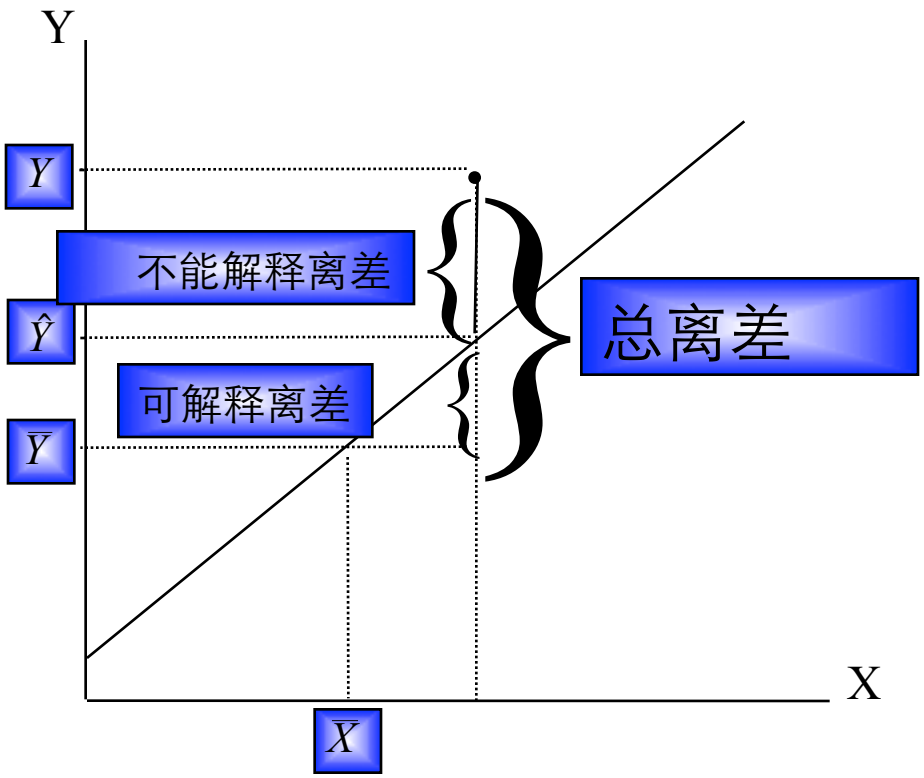
离差分解图

从图上看有：总离差 = 回归离差 + 剩余离差

$$y-\bar{y}=(\hat{y}-\bar{y})+(y-\hat{y})$$


剩余离差平方和
回归离差平方和
总离差平方和
$SSE = \sum (y - \hat{y})^2$
$SSR = \sum (\hat{y} - \bar{y})^2$
$SST = \sum (y - \bar{y})^2$
$y - \hat{y}$
$\hat{y} - \bar{y}$
$y - \bar{y}$
$\hat{y}$
$\bar{y}$
10
75
55
50
45
40
158 160 162 164 166 168 170 172 174 176 178
身高（X）

离差平方和的分解



$$(y-\bar{y}) = (y-\hat{y}) + (\hat{y}-\bar{y})$$

Total Deviation = Unexplained Deviation (Error) + Explained Deviation (Regression)

$$\sum(y-\bar{y})^2 = \sum(y-\hat{y})^2 + \sum(\hat{y}-\bar{y})^2$$

SST = SSE + SSR

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Percentage of total variation explained by the regression.

离差平方和的分解

- 1.总变差=回归变差+剩余变差

$$y-\bar{y}=(\hat{y}-\bar{y})+(y-\hat{y})$$

- 2. 两端平方后求和有：

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y})^2$$

总变差平方和（SST）　　回归平方和（SSR）　　残差平方和（SSE）

记为：SST = SSR + SSE　或　$L_{yy}=U+Q$

三个离差平方和的意义

1. 总（离差）平方和（*SST*、L_{yy}）
   - 反映因变量的 *n* 个观察值与其均值的总离差
2. 回归平方和（*SSR*、*U*）
   - 反映自变量 *x* 的变化对因变量 *y* 取值变化的影响，或者说，是由于*x*与*y*之间的线性关系引起的*y*的取值变化。
3. 残差平方和（*SSE*、*Q*）
   - 反映除 *x* 以外的其他因素对 *y* 取值的影响。

决定系数r^2

1. 判定系数=回归平方和占总离差平方和的比例

$$r^2 = \frac{SSR}{SST} = \frac{\text{回归平方和}}{\text{总离差平方和}} = 1 - \frac{\text{残差平方和}}{\text{总离差平方和}}$$

2. 判定系数=相关系数的平方，即$r^2=(r)^2$
3. 反映回归直线的拟合程度，衡量变量之间的相关程度。
4. 取值范围在 [0,1] 之间。

$r^2 \to 1$，说明回归方程拟合效果越好；

$r^2 \to 0$，说明回归方程拟合得越差。

*b*与*r*的关系：

$\begin{cases} r>0 \\ b>0 \end{cases}$　$\begin{cases} r<0 \\ b<0 \end{cases}$　$\begin{cases} r=0 \\ b=0 \end{cases}$

$$r = b\frac{S_x}{S_y} \quad ; \quad b = r\frac{S_y}{S_x}$$

判定系数与相关系数的区别：

- 判定系数无方向性，相关系数则有方向，其方向与样本回归系数 **b** 相同；
- 判定系数说明变量值的总离差平方和中可以用回归线来解释的比例，相关系数只说明两变量间关联程度及方向；

回归估计标准差与相关系数的关系

$$r = \sqrt{1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2}}$$

- 大样本条件下，近似地：

$$r = \sqrt{1 - \frac{\sum(y - \hat{y})^2 / n}{\sum(y - \bar{y})^2 / n}} = \sqrt{1 - \frac{S_e^{\,2}}{S_y^2}}$$

- 或：$S_e = S_y\sqrt{1 - r^2}$

估计标准差越小，则变量间相关程度越高，回归线对Y的解释程度越高。

§6.2　一元线性回归分析

一、回归分析概述

二、一元线性回归模型

三、回归估计标准差

★ 四、线性相关的显著性检验

五、回归估计与预测

线性相关的显著性检验

- 检验两个变量之间是否存在线性相关关系，可以从两个方面考虑：
  - 对回归系数 β 的显著性检验；
  - 对相关系数 r 的显著性检验；

- 对于一元线性相关而言，二者完全等价。

回归问题的方差分析 （F检验 ）

1. 提出假设。H_0： β=0（线性关系不显著）；

   H_1： β≠0（线性关系显著）

2. 确定检验统计量 $F = \dfrac{SSR/1}{SSE/(n-2)} \sim F(1, n-2)$

3. 确定显著性水平α，找出临界值F_α（ 1，n-2）

4. 计算统计量的值；

5. 作出决策：若$F \geq F_\alpha$,拒绝H_0; 若$F < F_\alpha$，接受H_0

例10.2 下面是20名工作人员的智商和某一次技术考试成绩,根据这个结果求出考试成绩对之上的回归方程。如果另有一名工作人员智商为120，则估计一下若让他也参加技术考试，将会的多少分?

工作人员	智商（x）	考试成绩（y）	回归值（y'）	工作人员	智商（x）	考试成绩（y）	回归值（y'）
1	89	55	57.86	11	84	53	54.2
2	97	74	63.7	12	121	82	81.2
3	126	87	84.87	13	97	58	63.7
4	87	60	56.4	14	101	60	66.6
5	119	71	79.76	15	92	67	60.1
6	101	54	66.6	16	110	80	73.2
7	130	90	87.8	17	128	85	86.3
8	115	73	76.8	18	111	73	73.9
9	108	67	71.7	19	99	71	65.2
10	105	70	69.5	20	120	90	80.5

解：经计算$\bar{x}=107,\bar{y}=71,\hat{\beta}_1=\dfrac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}=\dfrac{L_{xy}}{L_{xx}}=0.73,$

$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}=71-0.73\times107=-7.11,$

回归方程为$\hat{y}=0.73x-7.11,$

若$x=120,$则$\hat{y}=0.73\times120-7.11=80.5,$

若$x=97,$则$\hat{y}=0.73\times97-7.11=63.7,$

$$S^2=\sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}$$

$$=\sqrt{\frac{(55-57.86)^2+(74-63.7)^2+\ldots+(90-80.5)^2}{18}}$$

$$=6.27$$

$$\sum(x-\bar{x})^2=3748$$

$$S_{y_0}=6.27\times\sqrt{1+\frac{1}{20}+\frac{(97-107)^2}{3748}}=6.27\times1.037=6.5,$$

查t分布表，$t_{0.05/2}(18)=2.101,x=97$对应的$y_0$的.95的置信区间为:$63.7\pm2.101\times6.5,$

即　50.04～77.36

SPSS输出结果（二）

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.976[a]	.952	.949	2.4567

a. Predictors: (Constant), ÄÜÔ´Ïû°ÄÁ¿

方差分析表

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1676.440	1	1676.440	277.761	.000[a]
	Residual	84.498	14	6.036		
	Total	1760.938	15			

a. Predictors: (Constant), ÄÜÔ´Ïû°ÄÁ¿

b. Dependent Variable: ¹¤Òµ×Ü²úÖµ

样本相关系数r的显著性检验（t检验法）

目的 检验总体两变量间线性相关性是否显著

步骤

1.提出假设：

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

2.构造检验统计量：

$$t=r\sqrt{n-2}\Big/\sqrt{1-r^2}\sim t(n-2)$$

相关系数的显著性检验（t检验法）

步骤

3. 根据给定的显著性水平α，确定临界值 $t_{\alpha/2}$ ；

4. 确定原假设的拒绝规则：

若 $|t| < t_{\alpha/2}(n-2)$，则接受H_0，表示总体两变量间线性相关性不显著；

若 $|t| \geq t_{\alpha/2}(n-2)$，则拒绝$H_0$，表示总体两变量间线性相关性显著

5. 计算检验统计量并做出决策。

【例】检验工业总产值与能源消耗量之间的线性相关性是否显著　资料

解：已知$n=16, r=0.9757, \alpha=0.05$,则

提出假设：$H_0: \rho=0 \quad H_1: \rho\neq 0$

当 $H_0: \rho=0$ 成立时，则统计量

$$t=r\sqrt{n-2}\Big/\sqrt{1-r^2}\sim t(n-2)$$

有：$t=0.9757\sqrt{16-2}\Big/\sqrt{1-(0.9757)^2}=16.6616$

$\because t=16.6616\rangle t_{\alpha/2}(n-2)=t_{0.025}(14)=2.1448$

$\therefore$拒绝H_0，表示总体的两变量间线性相关性显著。

采用t检验方法

$$t_b = \frac{|b|}{SE_b} \qquad SE_b = \frac{s_{y.x}}{\sqrt{\sum (X - \overline{X})^2}}$$

其中$S_{y.x}$为各观察值Y到回归直线的距离的标准差，表示去除X影响后Y的变异程度.

§6.2　一元线性回归分析

一、回归分析概述

二、一元线性回归模型

三、回归估计标准差

四、线性相关的显著性检验

★ 五、回归估计与预测

回归方程的估计与预测

估计的前提：回归方程经过检验，证明 X 和 Y 的关系在统计上是显著相关的。

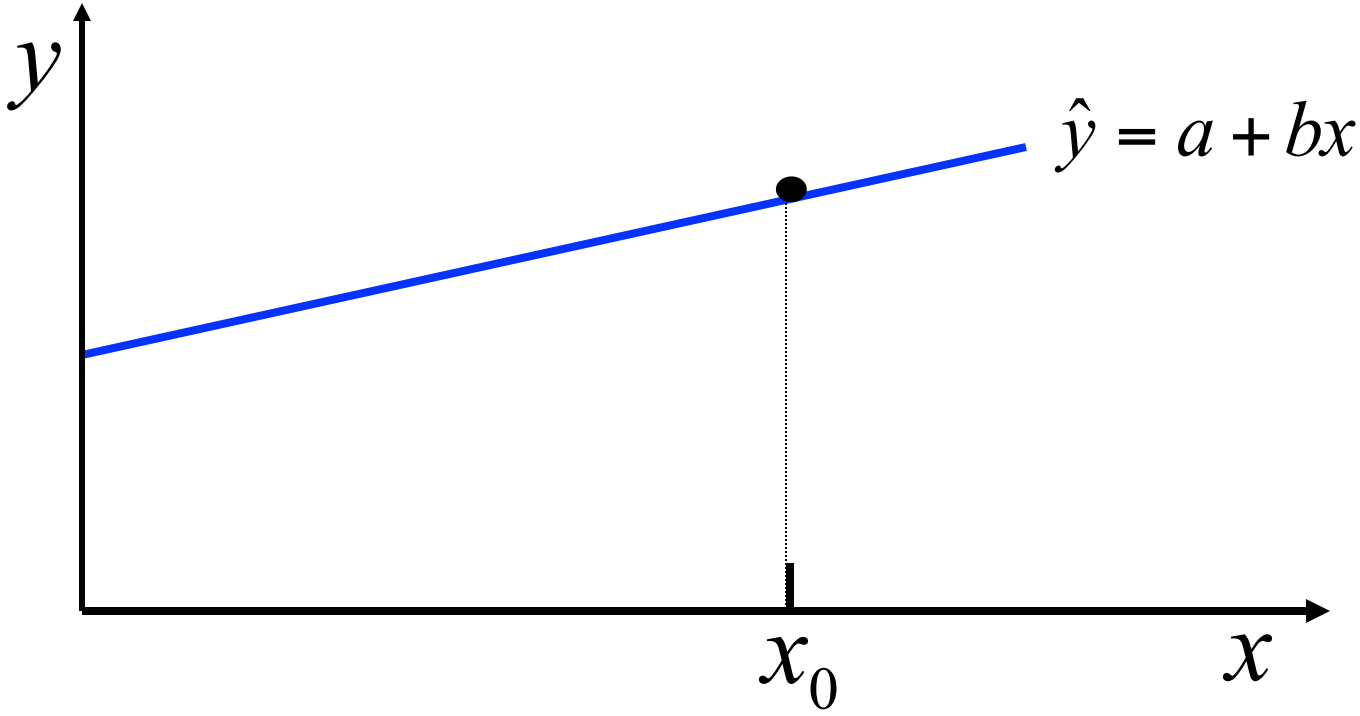
点估计

对于给定的 X 值，求出 Y 平均值的一个估计值或 Y 的一个个别值的预测值。

区间估计

对于给定的 X 值，求出 Y 的平均值的置信区间或 Y 的一个个别值的预测区间。

点估计

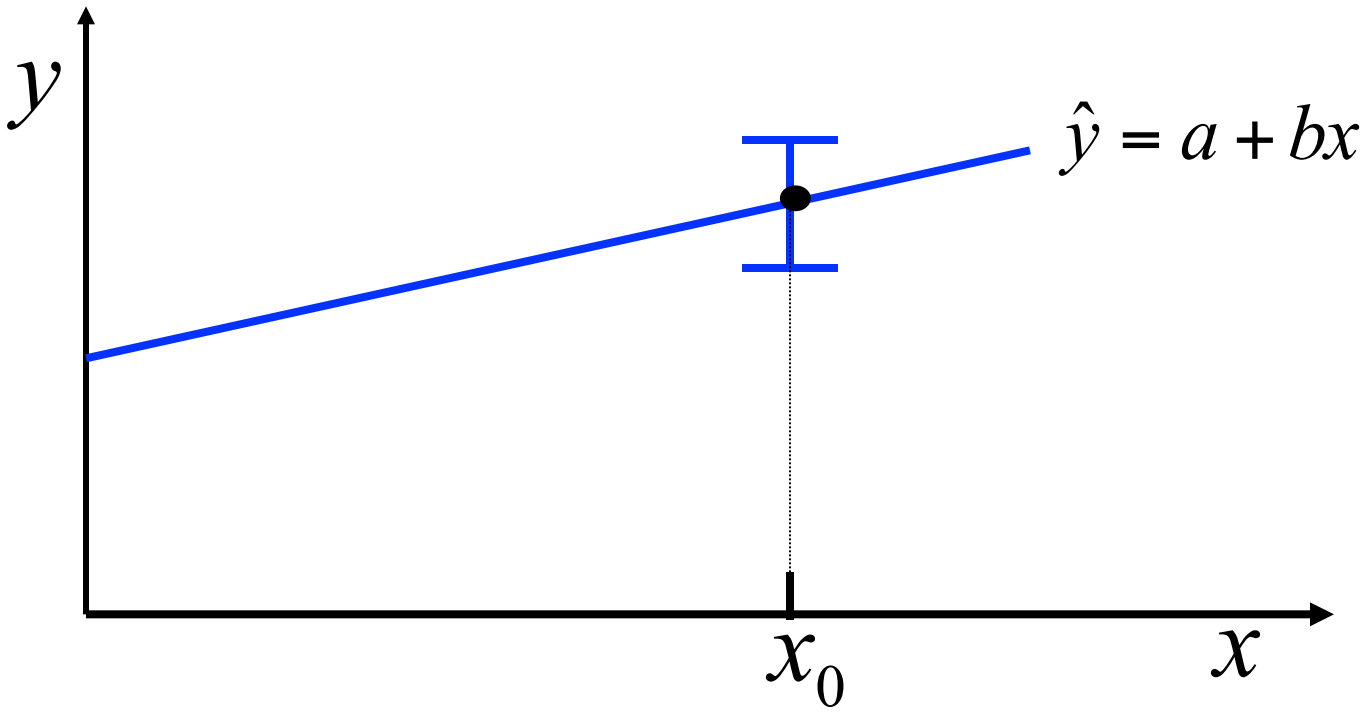


对于 $\hat{y} = -6.5142 + 0.7961x$

若 **x = 80**（十万吨），则：

$\hat{y} = -6.5142 + 0.7961 \times 80 = 57.1738$(亿元)

区间估计



对于给定的 $x = x_0$，Y 的1-α置信区间为：

$$\hat{y}_0 \pm t_{\alpha/2}\sigma_{\hat{y}}$$

自由度为n-2的 t 分布的 α 水平双侧分位数

预测的置信区间

估计值（均值）的置信区间：

$$\hat{Y}_0 \pm t_{\frac{\alpha}{2}}(n-2)\sqrt{MS_e\left[\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right]}$$

观测值的置信区间：

$$\hat{Y}_0 \pm t_{\frac{\alpha}{2}}(n-2)\sqrt{MS_e\left[1+\frac{1}{n}+\frac{(X-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right]}$$

预测区间

- 在1-α置信水平下，y_0的预测区间为：

$$\left(\hat{y}_0 \mp t_{(n-2)\alpha/2} S_e \sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right)$$

- 在大样本条件下，近似有：

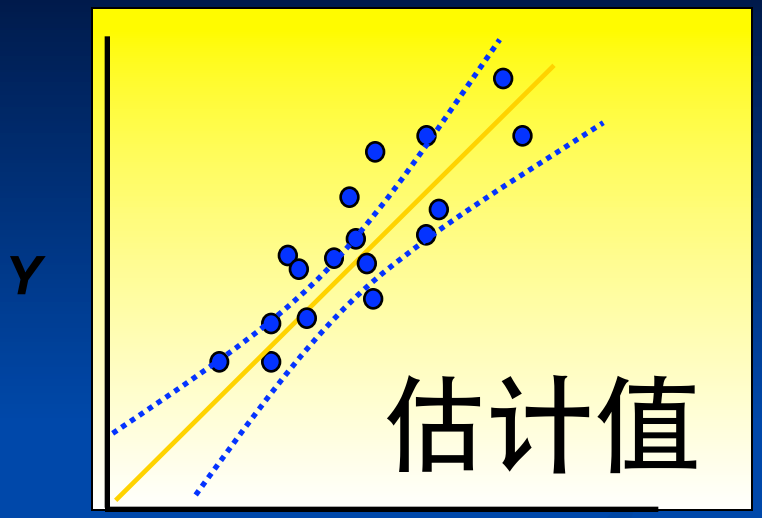
$$\hat{y}_0 \pm Z_{\alpha/2} S_e = \hat{y}_0 \pm Z_{\alpha/2}\sqrt{\frac{\sum (Y-\hat{Y})^2}{n-2}}$$

影响区间宽度的因素

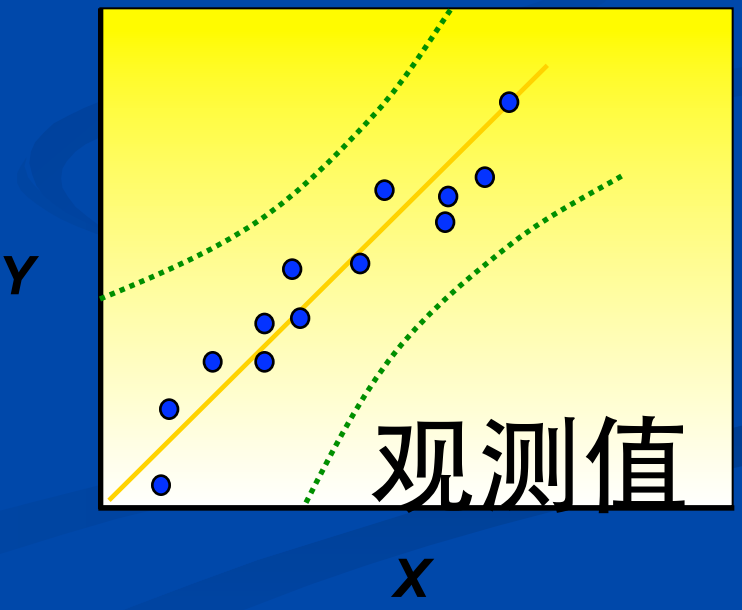
- 1. 置信水平 $(1-\alpha)$
  - 区间宽度随置信水平的增大而增大
- 2. 回归估计标准差 (Se)
  - 区间宽度随离散程度的增大而增大 ？
- 3. 样本容量
  - 区间宽度随样本容量的增大而减小 ？
- 4. 用于预测的x_0与$\bar{x}$的差异程度。
  - 区间宽度随x_0与$\bar{x}$的差异程度的增大而增大 ？

预测的置信区间

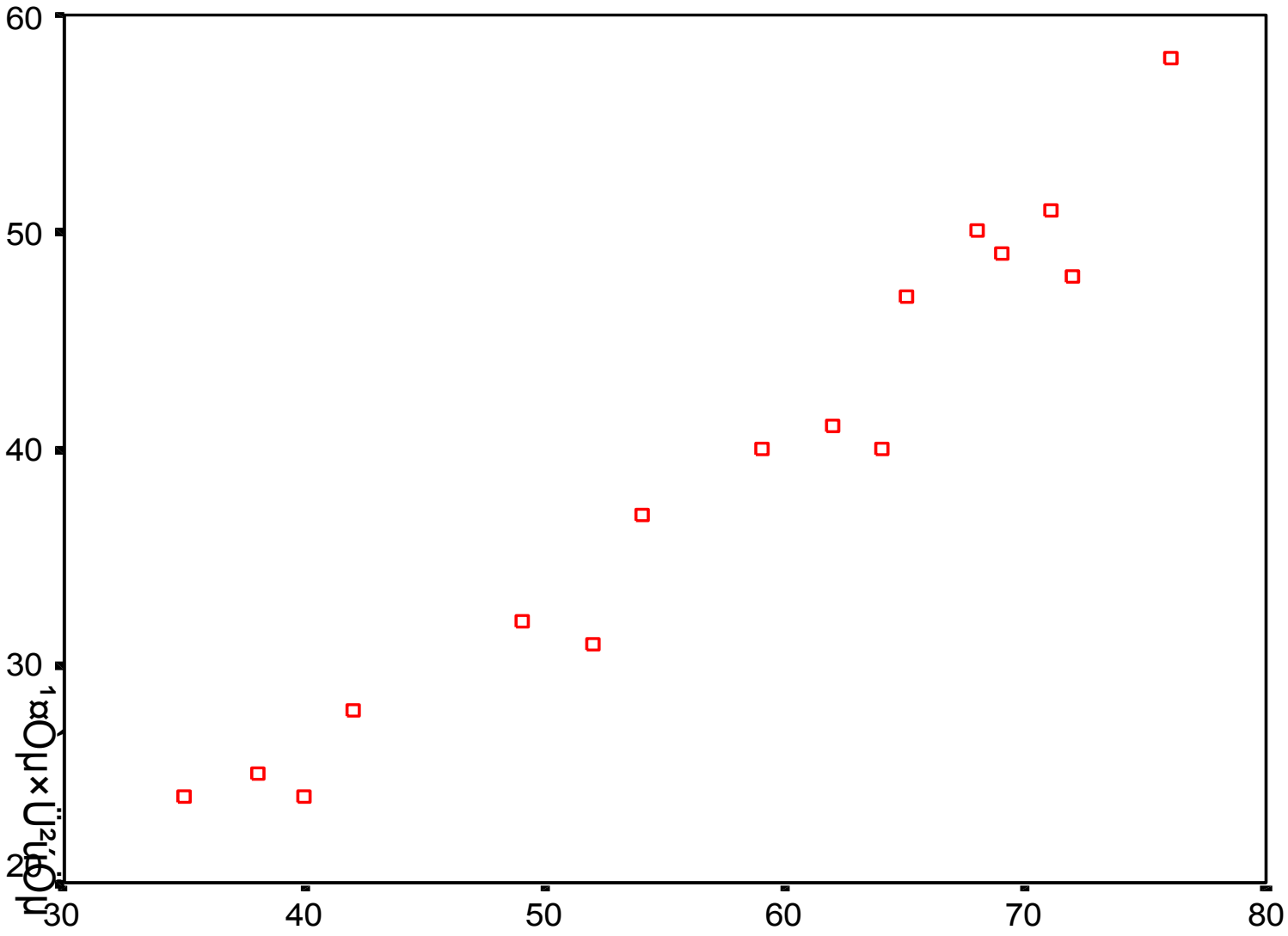
观测值的置信区间要比估计值的置信区间长；



- 预测点x_0距样本均值越远，置信区间越长，预测的准确性降低。



SPSS输出结果（一）



ÄÜÔ´Ïû°ÄÁ¿

SPSS输出结果（二）

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.976[a]	.952	.949	2.4567

a. Predictors: (Constant), ÄÜÔ´Ïû°ÄÁ¿

方差分析表

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1676.440	1	1676.440	277.761	.000[a]
	Residual	84.498	14	6.036		
	Total	1760.938	15			

a. Predictors: (Constant), ÄÜÔ´Ïû°ÄÁ¿

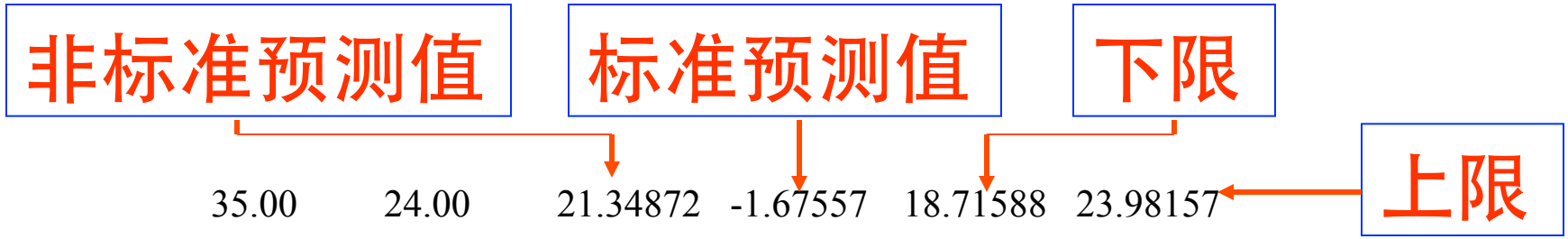
b. Dependent Variable: ¹¤Òµ×Ü²úÖµ

SPSS输出结果（三）

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-6.516	2.803		-2.325	.036
	ÄÜÔ´Ïû°ÄÁ¿	.796	.048	.976	16.666	.000

a. Dependent Variable: ¹¤Òµ×Ü²úÖµ



35.00	24.00	21.34872	-1.67557	18.71588	23.98157
38.00	25.00	23.73710	-1.44965	21.36539	26.10881
40.00	24.00	25.32935	-1.29904	23.12509	27.53360
42.00	28.00	26.92160	-1.14842	24.87796	28.96523
49.00	32.00	32.49447	-.62128	30.92932	34.05962
52.00	31.00	34.88284	-.39536	33.45997	36.30572
54.00	37.00	36.47509	-.24475	35.11637	37.83382
59.00	40.00	40.45572	.13179	39.12628	41.78516
62.00	41.00	42.84409	.35771	41.43978	44.24841
64.00	40.00	44.43634	.50832	42.94855	45.92413
65.00	47.00	45.23247	.58363	43.69437	46.77056
68.00	50.00	47.62084	.80955	45.90378	49.33791
69.00	49.00	48.41697	.88485	46.63245	50.20148
71.00	51.00	50.00922	1.03547	48.08053	51.93790
72.00	48.00	50.80534	1.11077	48.80060	52.81008
76.00	58.00	53.98984	1.41200	51.66055	56.31912

简单相关表

八个同类工业企业的月产量与生产费用

企业编号	月产量（千吨）X	生产费用（万元）Y
1	1.2	62
2	2.0	86
3	3.1	80
4	3.8	110
5	5.0	115
6	6.1	132
7	7.2	135
8	8.0	160

分组相关表

20个同类工业企业固定资产原值与平均每昼夜产量

平均每昼夜产量（吨）	固定资产原值（百万元）							
	35～40	40～45	45～50	50～55	55～60	60～65	65～70	$\sum f_Y$
600～650							1	1
550～600					1	2		3
500～550					2	1		3
450～500			1	5	1			7
400～450		2	2					4
350～400								0
300～350	2							2
$\sum f_X$	2	2	3	5	4	3	1	20

序号	能源消耗量（十万吨）x	工业总产值（亿元）y	x^2	y^2	xy
1	35	24	1225	576	840
2	38	25	1444	625	950
3	40	24	1600	576	960
4	42	28	1764	784	1176
5	49	32	2401	1024	1568
6	52	31	2704	961	1612
7	54	37	2916	1369	1998
8	59	40	3481	1600	2360
9	62	41	3844	1681	2542
10	64	40	4096	1600	2560
11	65	47	4225	2209	3055
12	68	50	4624	2500	3400
13	69	49	4761	2401	3381
14	71	51	5041	2601	3621
15	72	48	5184	2304	3456
16	76	58	5776	3364	4408
合计	916	625	55086	26175	37887

例　在研究我国人均消费水平的问题中，把人均国民收入记为y,把人均国民收入记为x。我们收集到1981--1993年13年的样本数据（x_i,y_i）,i=1,2,…,13。数据如下表:

年份	人均国民收入	人均消费金额	年份	人均国民收入	人均消费金额
1981	393.8	249	1988	1068.8	643
1982	419.14	267	1989	1169.2	699
1983	460.86	289	1990	1250.7	713
1984	544.11	329	1991	1429.5	803
1985	668.29	406	1992	1725.9	947
1986	737.73	451	1993	2099.5	1148
1987	859.97	513			

$$\overline{X} = \frac{1}{13}\sum_{i=1}^{13} x_i = 986.733$$

$$\overline{Y} = \frac{1}{13}\sum_{i=1}^{13} y_i = 573.615$$

$$\hat{\beta}_1 = \frac{L_{xy}}{L_{xx}} = 0.52638$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}\overline{x} = 54.219$$

回归方程为：$\hat{Y} = 54.219 + 0.526X$

应用多元回归方程时存在的潜在问题

1、共线性

$x_1,x_2,\ldots,x_p$之间存在密切的线性关系，称它们之间存在着多重共线性（Multi-collinearity）。此时对回归系数的估计不稳定。

2、因果关系

回归分析能表现出变量彼此关联或有联系，但不能证明其因果关系。要确定x与y存在因果关系，必须有很强的逻辑性或理论性的基础。即使有很强的逻辑性和统计相关性，也只是表明可能存在因果关系。

3、回归系数的大小

只有在计量单位相同或数据标准化的情况下，与各自变量相联系的回归系数大小才能直接比较。

4、样本容量

R^2受n对于自变量个数p的影响。一般观测数n至少等于自变量个数p的10 – 15倍。

回归分析与相关分析的区别

1、在回归分析中，变量y称为因变量，处于被解释的地位。而在相关分析中，x与y处于平等地位；

2、相关分析中，x,y全是随机变量，而在回归分析中，因变量y是随机变量，自变量x可以是随机变量，也可以是非随机的。通常回归模型中假定x是非随机的精确变量；

3、相关分析的研究是为了刻画两变量间线性相关的密切程度。而回归分析不仅可以揭示x对y的影响大小，还可以由回归方程进行预测和控制。

回归分析实例

例10.3 中国民航客运量的回归模型。为了研究我国民航客运量变化趋势及其成因，以民航客运量（万人）作为因变量y, 以国民收入x_1(亿元)、消费额x2(亿元)、铁路客运量x3(万人)、民航航线里程x4(万公里)、来华旅游入境人数x5（万人）为影响民航客运量的主要因素。

例 某个公司营销经理认为:决定目标消费者为接受广域服务(PCS服务的品牌)支付的价格有5个变量：覆盖面、移动性、音质、远距离接听和打出电话的能力、月均缴费。

6个变量（1个因变量和5个自变量）的数据由调查获得。5个自变量以9级评分制测量。

“9”表示某“特性”很重要；

“1”表示该“特性”很不重要。

提出多元回归模型:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 + b_5 x_5$$

$\hat{y}$---因变量，每月接受广域服务所愿支付的价格；

b_0---常数项或y轴截距；

b_1—b_5---回归系数；

x_1---覆盖范围的重要性评分；

x_2---移动性的重要性评分；

x_3---音质的重要性评分；

x_4---远距离接收和打出电话能力的重要性评分；

x_5---月均电话费的重要性评分。

通过统计软件分析，得到回归方程如下：

$$\hat{y} = 0.82 + 0.44 x_1 + 0.69 x_2 + 0.21 x_3 + 0.45 x_4 + 1.44 x_5$$

b_1—b_5均为正，表明5个变量的重要性评分越高，人们对广域服务愿意支付的价格越高。

回归分析结果:

$R = 0.857$，$R^2 = 0.7495$， 调整$R^2 = 0.743$

$F（5，194）= 116.09$，$p<0.001$, 估计标准差：1.4863

$p<0.001$表明回归方程有意义。所有自变量整体对因变量具有预测作用，它们之间具有线性关系。

$R^2 = 0.743$表明消费者愿意支付价格的变异中有74.3％可以被5个自变量或预测变量的变异所解释。

	回归系数及其显著性检验结果					
	B	B的标准差	BETA	BETA的标准差	t(194)	p值
截距	0.82	1.67			0.49	0.62
覆盖面	0.44	0.10	0.21	0.05	4.25	0.00
移动性	0.69	0.07	0.52	0.05	10.54	0.00
音质	0.21	0.13	0.07	0.04	1.67	0.10
远距离收发能力	0.45	0.12	0.21	0.06	3.65	0.00
平均月租费	1.44	0.17	0.32	0.04	8.33	0.00

自变量对因变量的影响除音质不显著外，其余都显著。从BETA一列可以看出，影响大小依次为：移动性、平均月缴费、覆盖面、远距离收发能力。最不重要的是音质。