一元线性回归

——最小二乘估计

主讲：黎雅谨

变量与变量之间的关系?

一、确定性关系

当一个变量给定时，就确定另一个变量的值与之对应。

某种商品的销售额(y)与销售量(x)之间的函数关系：

$$y = p\,x \text{ (p 为单价)}$$

圆的面积(S)与半径(R)之间的函数关系：

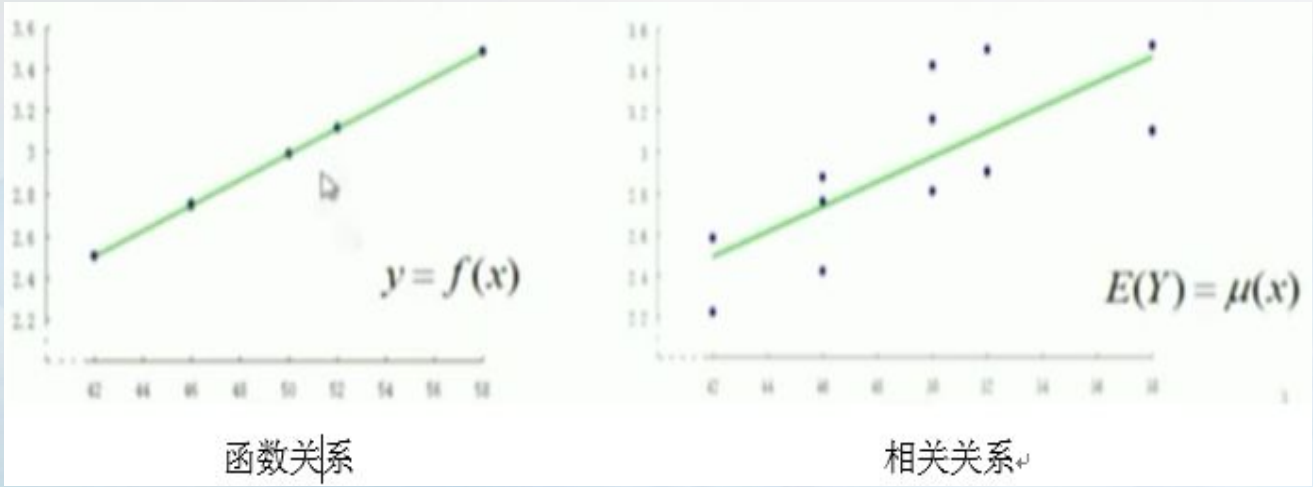
$$S = \pi R^2$$

身高与体重的关系？

人的脚掌的长度与身高的关系？

二、相关性关系

当一个变量给定时，受影响的另一个变量的值不能完全确定，而是在一定范围内变化。



函数关系

相关关系

我们以一个例子来建立一元线性回归模型：

例1 为研究某市居民家庭人均年消费支出Y(百元)与人均年可支配收入X(百元)之间的关系,2011年中国统计年鉴上收集了1985-2010年26年的统计数据。

年	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997
人均可支配收入x(百元)	8.12	9.84	11.09	12.78	14.49	16.91	18.92	21.95	27.1	36.34	43.75	50.23	53.02
人均消费性支出y(百元)	7.11	8.94	10.44	13.23	13.83	15.70	17.54	19.29	23.97	31.27	40.52	44.67	49.20
年	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
人均可支配收入x(百元)	54.43	58.28	61.76	65.72	72.38	80.94	92.21	102.44	115.70	137.15	157.09	171.91	191.00
人均消费性支出y(百元)	49.57	53.77	54.72	57.25	63.60	71.18	79.73	86.23	93.99	108.76	122.69	135.07	147.55

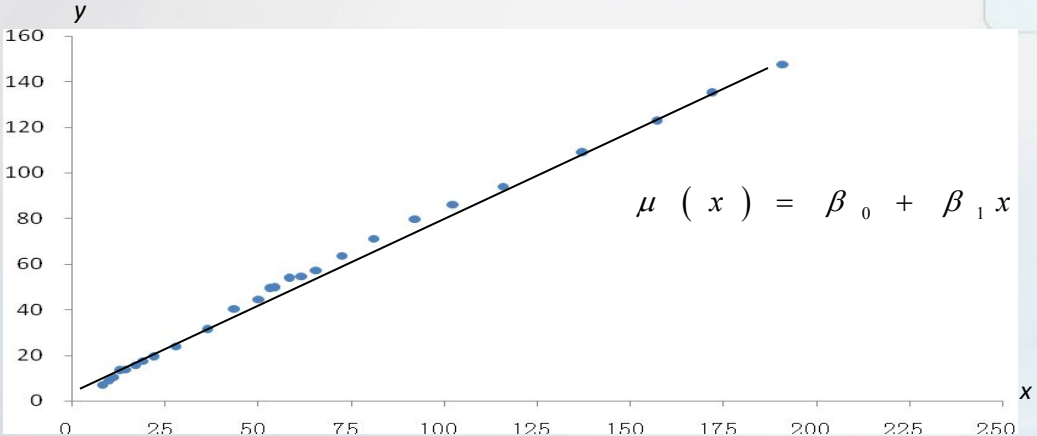


图1　　城镇居民人均年消费支出*y*与收入*x*之间的数据散点图

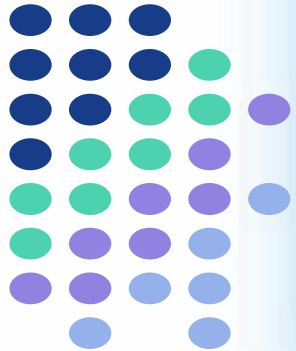
$$y = \beta_0 + \beta_1 x + \varepsilon$$

对从总体(x,Y)中抽取的一个样本

$(x_1,Y_1),(x_2,Y_2),\cdots,(x_n,Y_n)$

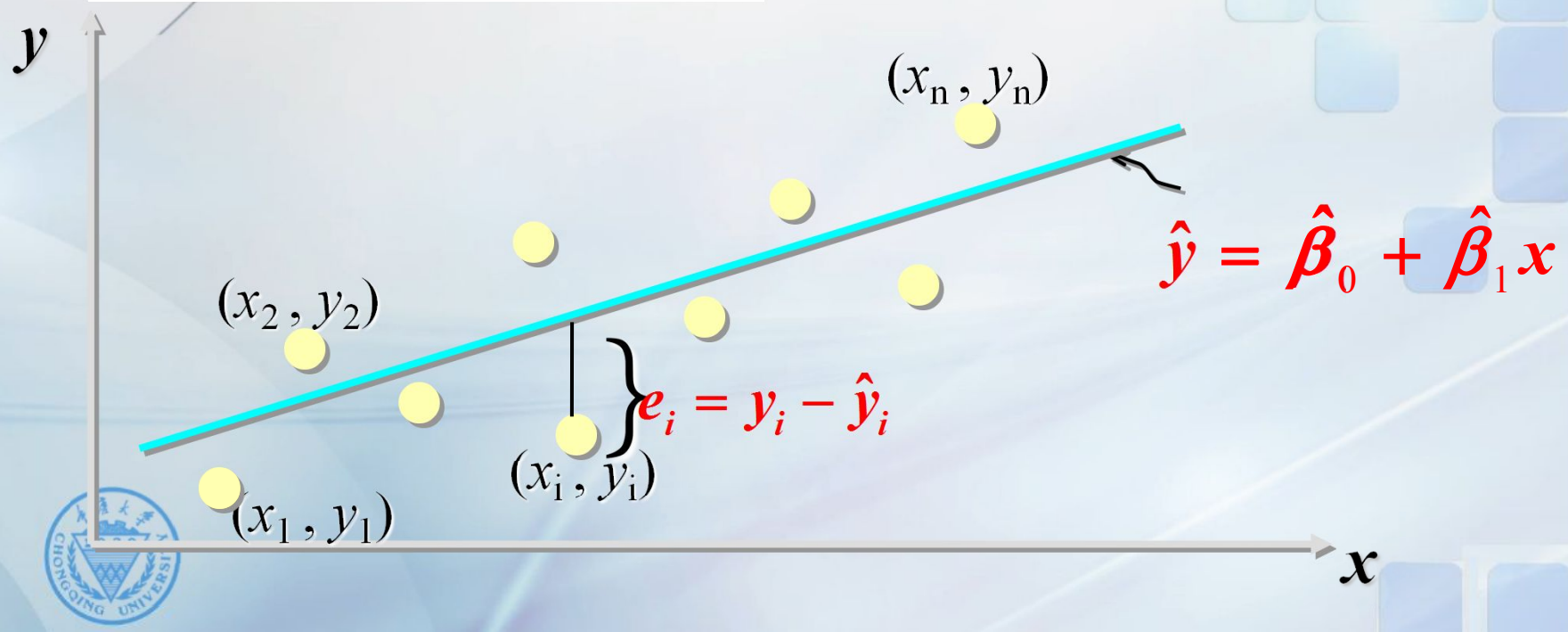
一元线性回归模型：

$$\begin{cases} Y_i=\beta_0+\beta_1 x_i+\varepsilon_i, & i=1,2,\ldots,n \\ \varepsilon_i\sim N(0,\sigma^2), & i=1,2,\ldots,n \\ Cov(\varepsilon_i,\varepsilon_j)=0, & i\neq j,i,j=1,2,\ldots,n \end{cases}$$



给定样本观测值为$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$，

参数 β_0 和 β_1 的估计？



最小二乘法思想

$$Q(\beta_0,\beta_1)=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2=\sum_{i=1}^{n}\varepsilon_i^2$$

$$Q(\hat{\beta}_0,\hat{\beta}_1)=\min Q(\beta_0,\beta_1)$$

最小二乘法(求解)

要求 $\min Q(\beta_0, \beta_1)$

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

$$\begin{cases} \left.\dfrac{\partial Q}{\partial \beta_0}\right|_{\beta_0=\hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \left.\dfrac{\partial Q}{\partial \beta_1}\right|_{\beta_1=\hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^{n} e_i = 0 \\ \sum_{i=1}^{n} x_i e_i = 0 \end{cases}$$

$$\begin{cases} n\hat{\beta}_0 + n\bar{x}\hat{\beta}_1 = n\bar{y} \\ n\bar{x}\hat{\beta}_0 + \left(\sum_{i=1}^{n} x_i^2\right)\hat{\beta}_1 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

称为正则方程。

其中 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

系数行列式

$$D = \begin{vmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^{n} x_i^2 \end{vmatrix} = n\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = n\sum_{i=1}^{n}(x_i - \bar{x})^2$$

由于x_i不全相等，所以D≠0,方程组有唯一解。

$$\begin{cases} \hat{\beta}_1 = \dfrac{L_{xy}}{L_{xx}} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$ ---最小二乘估计

其中 $L_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2$

$$L_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}$$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ ---回归方程

- 对于例1，应用EXCEL，**在Excel平台上计算，工具栏里有“数据分析”，含有“回归”，按步骤输入数据。**

计算结果(SUMMARY OUTPUT)

回归统计	
Multiple R	0.997635
R Square	0.995276
Adjusted R Square	0.995079
标准误差	2.8707
观测值	26

方差分析

	df	SS	MS	F	Significance F
回归分析	1	41666.55	41666.55	5056.056	2.0082E-29
残差	24	197.7821	8.24092		
总计	25	41864.33			

	Coefficients	标准误差	t Stat	P-value	下限 95.0%	上限 95.0%
Intercept	4.551596	0.901403	5.049459	3.67E-05	2.691192	6.412
X Variable 1	0.771799	0.010854	71.10595	2.0082E-29	0.749397	0.794201

$$\hat{y} = 4.5516 + 0.7718x$$