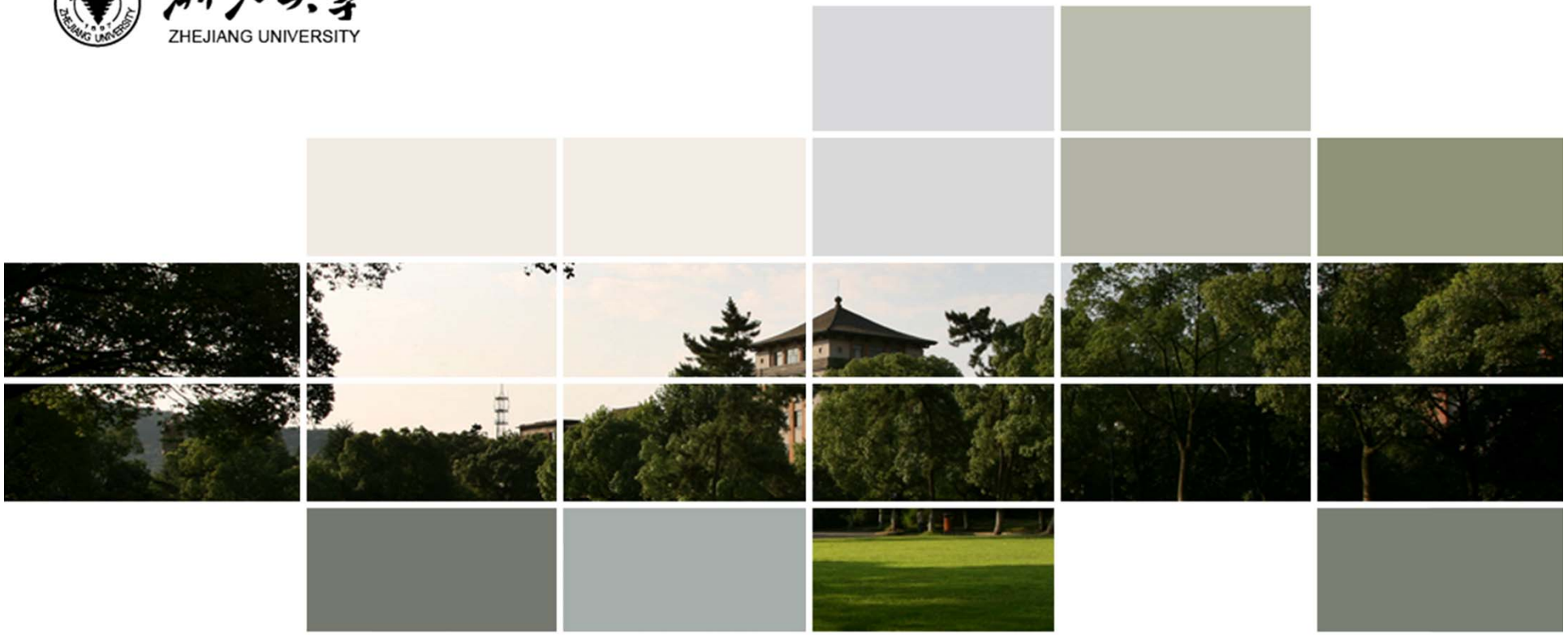
浙江大学
ZHEJIANG UNIVERSITY



第63讲　一元线性回归分析

变量与变量之间的关系 $\begin{cases} \text{确定性关系} \\ \text{相关性关系} \end{cases}$

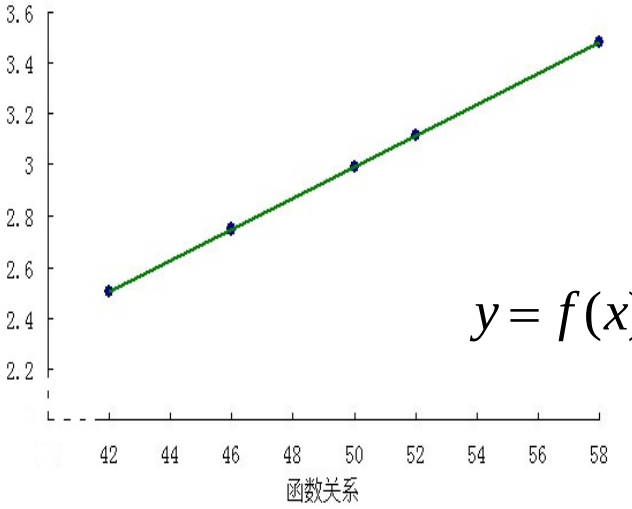
一、确定性关系：

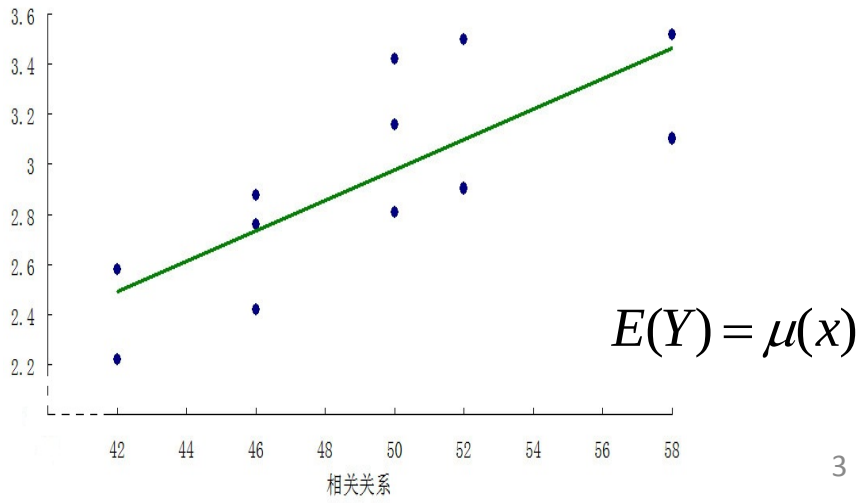
当自变量给定一个值时，就确定应变量的值与之对应。如：在自由落体中，物体下落的高度h与下落时间t之间有函数关系：

$$h = \frac{1}{2} g t^2$$

二、相关性关系：

变量之间的关系并不确定，而是表现为具有随机性的一种“趋势”。即对自变量x的同一值，在不同的观测中，因变量Y可以取不同的值，而且取值是随机的，但对应x在一定范围的不同值，对Y进行观测时，可以观察到Y随x的变化而呈现有一定趋势的变化。





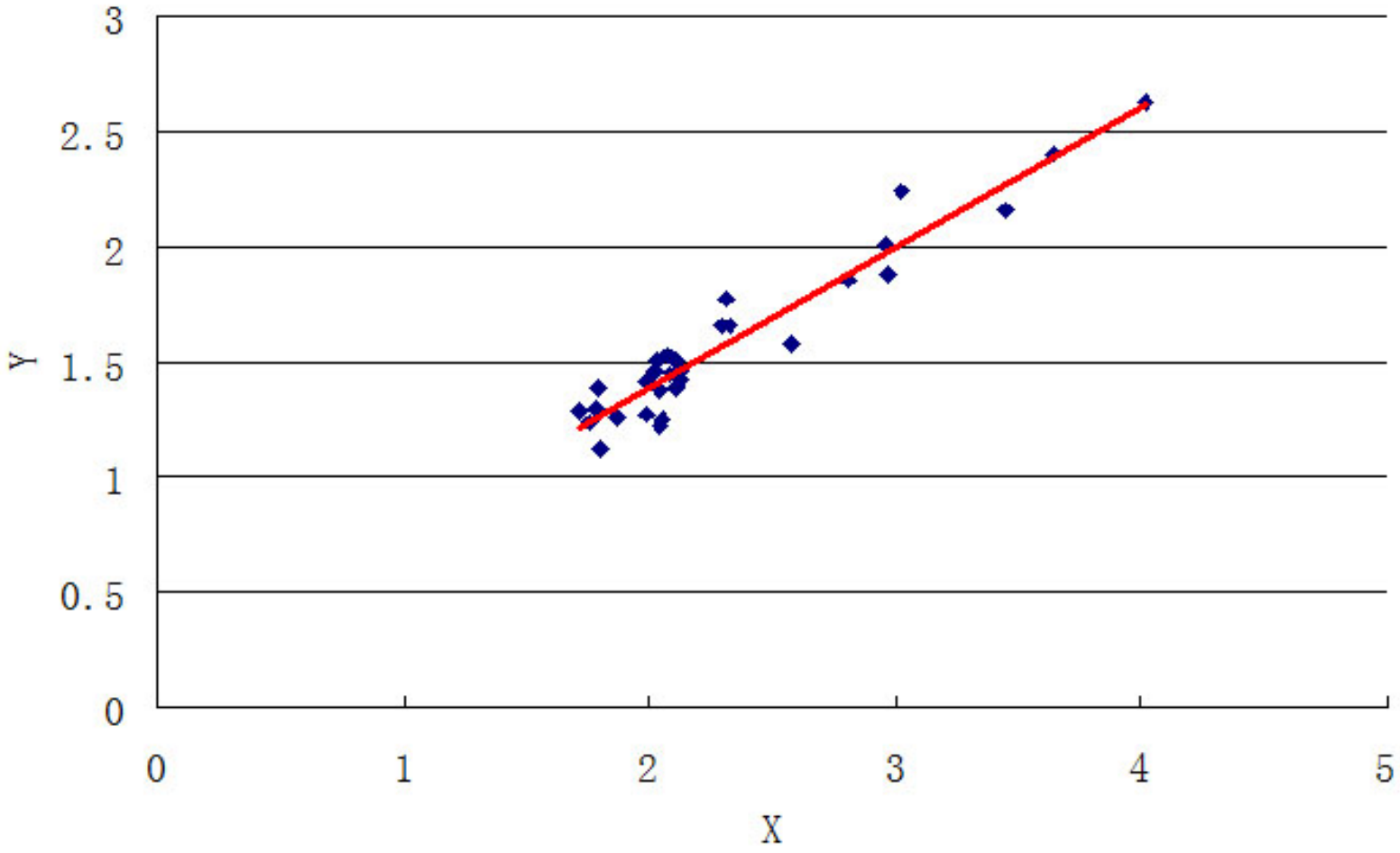
- 如：身高与体重，不存在这样的函数可以由身高计算出体重，但从统计意义上来说，身高者，体也重。
- 如：父亲的身高与儿子的身高之间也有一定联系，通常父亲高，儿子也高。

我们以一个例子来建立回归模型

- 例1：根据2013年《中国统计年鉴》的数据，2012年中国各地区城镇居民人均年消费支出和可支配收入数据见下表.

地区	可支配收入x(万元)	消费支出y(万元)	地区	可支配收入x(万元)	消费支出y(万元)
北京	3.647	2.405	上海	4.019	2.625
天津	2.963	2.002	江苏	2.968	1.883
河北	2.054	1.253	浙江	3.455	2.155
山西	2.041	1.221	安徽	2.102	1.501
内蒙古	2.315	1.772	福建	2.806	1.859
辽宁	2.322	1.659	江西	1.986	1.278
吉林	2.021	1.461	山东	2.576	1.578
黑龙江	1.776	1.298	河南	2.044	1.373

地区	可支配收入x(万元)	消费支出y(万元)	地区	可支配收入x(万元)	消费支出y(万元)
湖北	2.084	1.450	云南	2.107	1.388
湖南	2.132	1.461	西藏	1.803	1.118
广东	3.023	2.240	陕西	2.073	1.533
广西	2.124	1.424	甘肃	1.716	1.285
海南	2.092	1.446	青海	1.757	1.235
重庆	2.297	1.657	宁夏	1.983	1.407
四川	2.031	1.505	新疆	1.792	1.389
贵州	1.870	1.259			



散点图　X:可支配收入，
Y:消费支出

可支配收入x的变化是引起消费支出Y的变化的主要因素，其他因素的影响是次要的.

Y称为**响应变量**，x称为**解释变量**.

从散点图看出，引起消费支出Y的变化的主要部分可以表示为

$\mu(x)=\beta_0+\beta_1 x$，其中β_0,β_1是未知参数.

另一部分是由其他随机因素引起的，记为ε

即$Y=\beta_0+\beta_1 x+\varepsilon$.

对从总体(x,Y)中抽取的一个样本

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

一元线性回归模型：

$$\begin{cases} Y_i=\beta_0+\beta_1 x_i+\varepsilon_i, i=1,2,\ldots,n, \\ \varepsilon_i \sim N\left(0,\ \sigma^2\right),\text{且相互独立}, \\ \beta_0,\beta_1(\text{回归系数}),\sigma^2\text{未知}. \end{cases}$$

样本值为$(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$.

根据样本

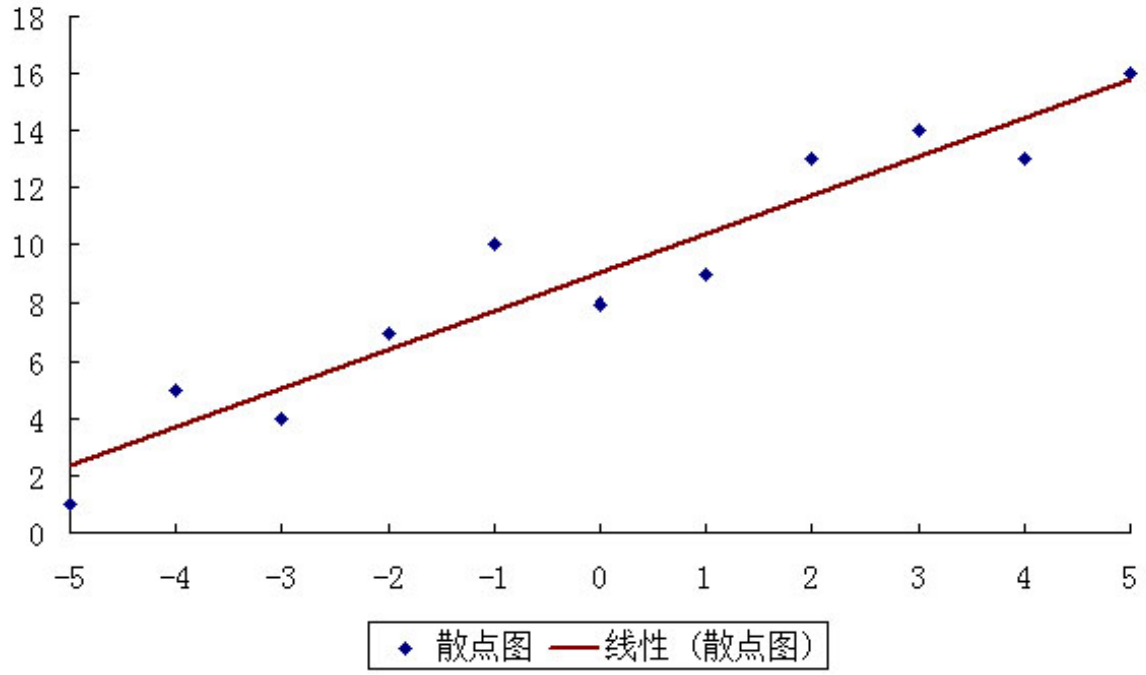
估计β_0，β_1，

记为$\hat{\beta}_0$，$\hat{\beta}_1$，

称为y关于x

一元线性回归

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$



一元线性回归要解决的问题：

参数估计：
- (1) β_0, β_1的估计；
- (2) σ^2的估计；

参数检验及模型应用：
- (3) 线性假设的显著性检验；
- (4) 回归系数β_1的置信区间；
- (5) Y的点预测.

(1)β_0,β_1的估计(采用最小二乘法)

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

求估计$\hat{\beta}_0,\hat{\beta}_1$,使$Q(\hat{\beta}_0,\hat{\beta}_1)=\min_{\alpha,\beta}Q(\beta_0,\beta_1)$.

$$\left.\frac{\partial Q}{\partial\beta_0}\right|_{\substack{\beta_0=\hat{\beta}_0\\ \beta_1=\hat{\beta}_1}}=0,\quad \left.\frac{\partial Q}{\partial\beta_1}\right|_{\substack{\beta_0=\hat{\beta}_0\\ \beta_1=\hat{\beta}_1}}=0.$$

$$\frac{\partial Q}{\partial\beta_0}=-2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)=0,$$

$$\frac{\partial Q}{\partial\beta_1}=-2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)x_i=0.$$

整理得正规方程系数行列式

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

$$(\sum_{i=1}^{n}x_i)\beta_0+(\sum_{i=1}^{n}x_i^{\,2})\beta_1=\sum_{i=1}^{n}x_iy_i.$$

$$\hat{\beta}_0+\bar{x}\hat{\beta}_1=\bar{y},$$

$$s_{xx}\hat{\beta}_1=s_{xy}.$$

记号: $\bar{y}=\frac{1}{n}\sum_i y_i,\ \bar{x}=\frac{1}{n}\sum_i x_i,\ s_{xx}=\sum_i(x_i-\bar{x})^2,$

$$s_{xy}=\sum_i(x_i-\bar{x})(y_i-\bar{y}),\ s_{yy}=\sum_i(y_i-\bar{y})^2.$$

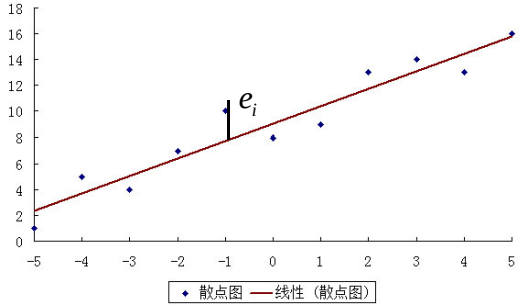
β_0,β_1的最小二乘估计:

$$\hat{\beta}_0=\bar{y}-\bar{x}\hat{\beta}_1,$$

$$\hat{\beta}_1=s_{xy}/s_{xx}.$$

(2) σ^2 的估计

记 $e_i = y_i - \hat{y}_i$ ——残差，e_i 是 ε_i 的估计



注意到 $\sigma^2 = D(\varepsilon_i) = E(\varepsilon_i)^2$,

用残差平方和 $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ 估计 σ^2.

可以证明, $E\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right) = (n-2)\sigma^2$

因此，$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ 是 σ^2 的无偏估计.

- 对于例1，应用EXCEL计算（具体结果见实验27），所得回归方程及方差估计分别为：

回归方程：$\hat{y} = 0.1707 + 0.6089x.$

方差σ^2估计$s^2 = 0.010736.$

方差分析表

	自由度	平方和	均方	F值	P_值
回归	1	3.800452	3.800452	353.987	8.54E-18
误差	29	0.311348	0.010736		显著！
总的	30	4.1118			

S^2

	Coef.	标准误差	t Stat	P value	Lower 95%	Upper 95%
Intercept	0.1707	0.0774	2.2046	0.0356	0.012	0.329
X	0.6089	0.0324	18.815	8.54E-18	0.543	0.675

$\hat{\beta}_0$ $\hat{\beta}_1$