应用数理统计

Ch5 回归分析

2014年5月5日

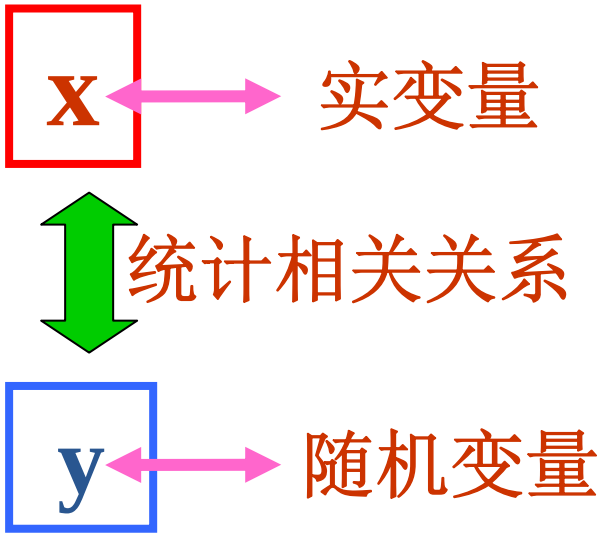
CH5 回归分析

- 回归分析基本概念
- 一元线性回归
- 多元线性回归
- 非线性回归
- 逐步回归分析

5.1 回归分析基本概念

变量间的关系

- 确定性关系或函数关系y=f(x)
- 非确定性关系
  - 人的身高和体重
  - 家庭的收入和消费
  - 粮食的施肥量和产量
  - 学生缺课次数和考试成绩
  - ……



统计相关关系： 变量y的取值与变量x有关，对给定的x值，变量y 是一个与x有关的随机变量

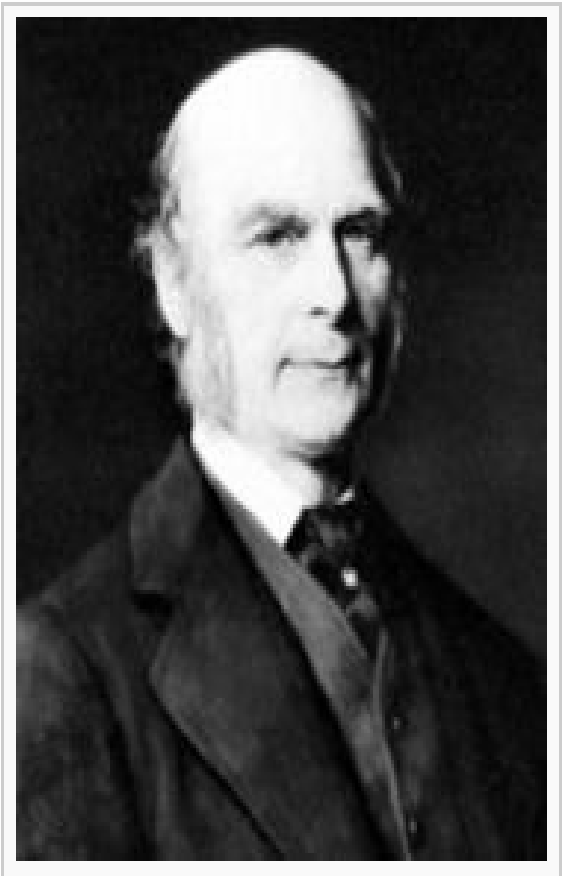
回归分析：就是研究变量间的统计相关关系的一种统计分析的方法.

根据变元的统计数据,用一个函数来近似表示变元间的统计相关关系,这个函数叫回归方程或回归函数

回归分析方法的分类:

- 线性回归
  - 一元回归
  - 多元回归
- 非线性回归
  - 多元回归

法兰西斯·高尔顿（Francis Galton，英国,1822－1911）



“维多利亚女王时代最博学的人”

优生学家、人类学家、探险家、
地理学家、发明家、气象学家、
统计学家、心理学家、 遗传学家

“种族主义者和法西斯的鼻祖和精神领袖”

统计学上的贡献:

首次提出相关性概念

建立回归分析的方法

1886年高尔顿与皮尔逊合作，收集分析了1078对父亲和儿子的身高数据，发表<<遗传中向平均身高的回归现象>>的论文

父身高 子身高

(x_i, y_i)，$i = 1, 2, \cdots, 1078$

得到直线的方程为

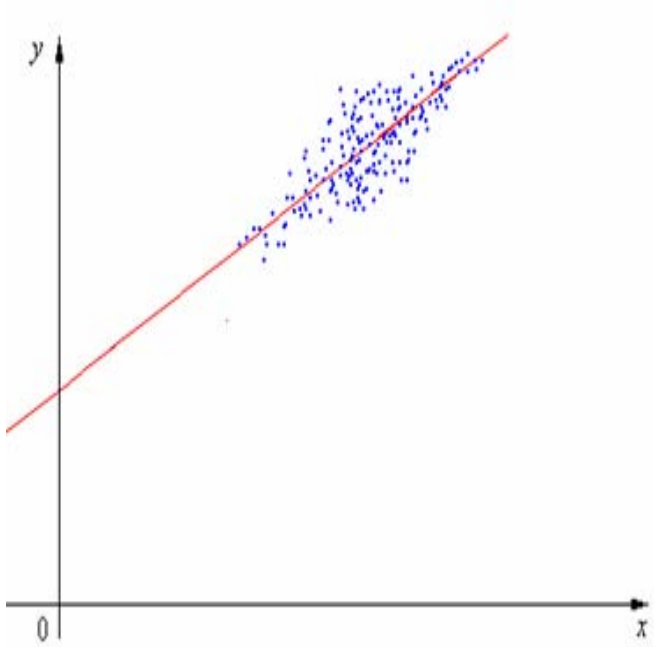
$\hat{y} = 0.8567 + 0.516x$ （单位：米）

例：$x = 1.900 \rightarrow \hat{y} = 1.837$

$x = 1.837 \rightarrow \hat{y} = 1.805$

$x = 1.600 \rightarrow \hat{y} = 1.682$

$x = 1.682 \rightarrow \hat{y} = 1.725$

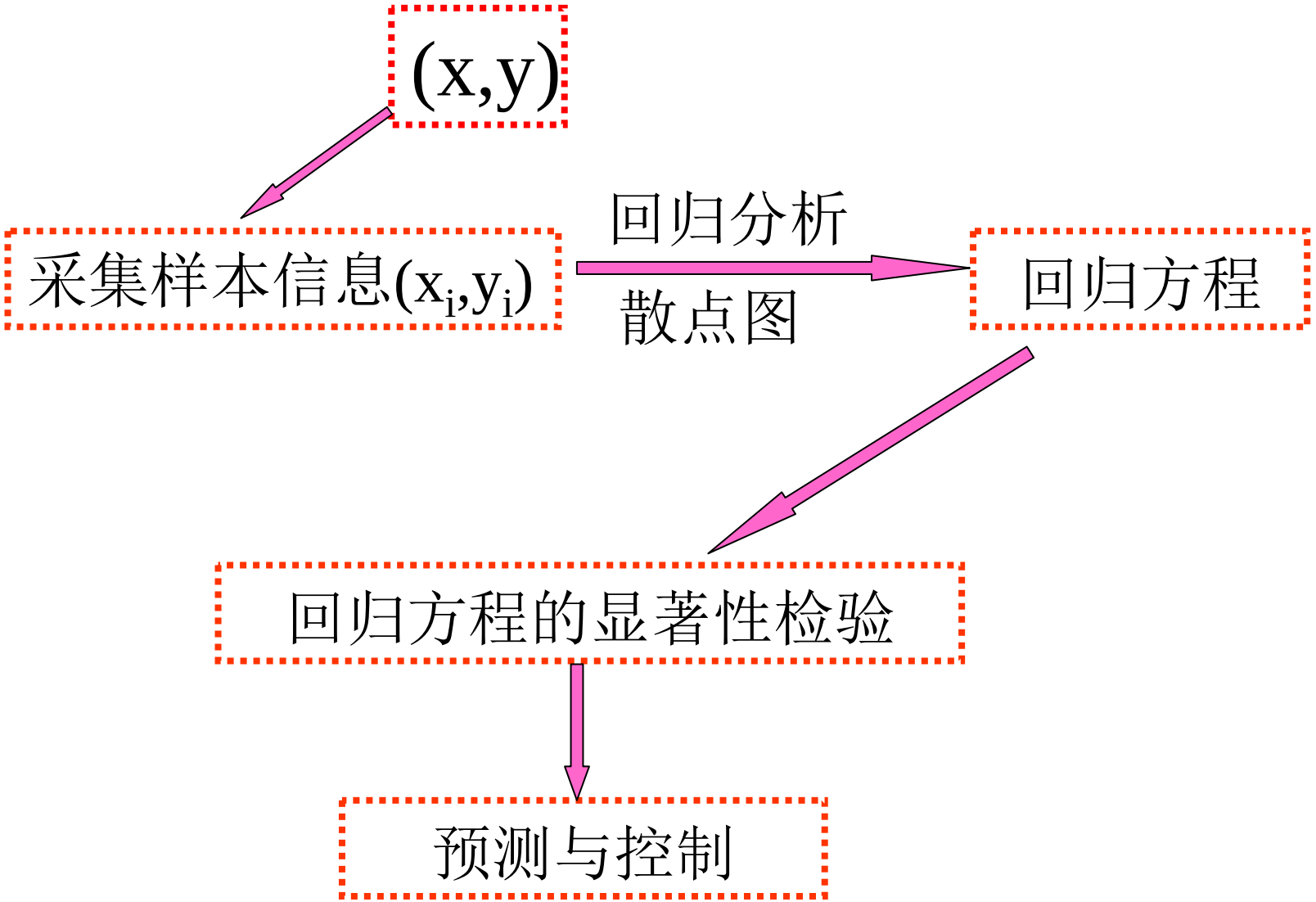


本例中：

统计相关关系：父身高与子身高

回归方程：　y=0.8567+0.516x

回归分析基本步骤:



5.2 一元线性回归

一元线性回归的模型:

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

x为确定性变量,也称解释变量或自变量;

y为被解释变量,响应变量,因变量

β_0和 β_1为未知的待估计参数

ε 为误差项, 它表示变元x与y间不能用线性关系解释的因素

根据变元（X，Y）的一组观测值(x_i, y_i)，$(i=1,2,\ldots,n)$代入上述一元线性回归模型，得：

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (i = 1,2,...,n)$$

随机向ε_i应满足一下三个前提条件：

1) 正态性：　　$\varepsilon_i \sim N(0,\sigma^2)$

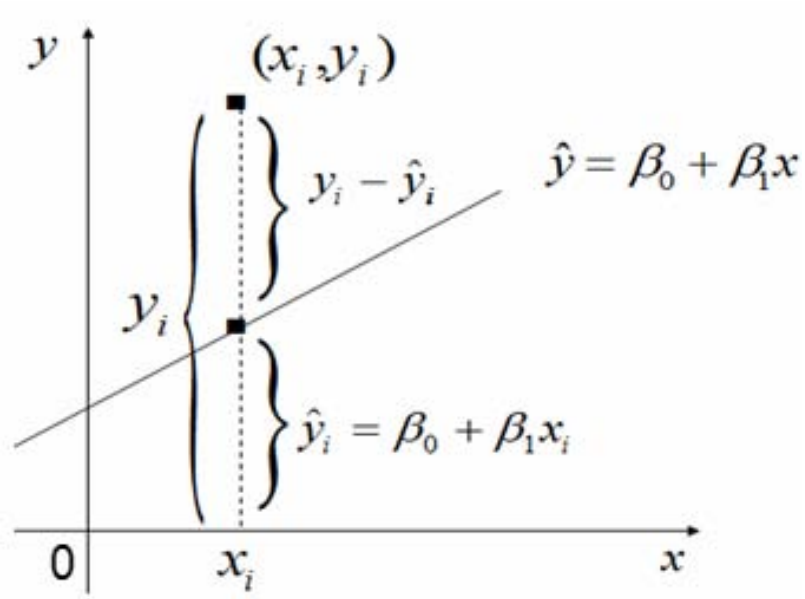
2) 独立性：　　ε_i相互独立

3）方差齐性：　ε_i的方差相同与i无关

求回归方程的问题

已知 (X,Y)的一组观测值 $(x_i, y_i)\ (i = 1, 2, ..., n)$,

求一条与所有观测值点最接近的直线 $\hat{y} = \beta_0 + \beta_1 x$



等价于求参数β_0和β_1,使得

$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 \text{ 取极小值}$$

求$Q=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}[y_i-(\beta_0+\beta_1 x_i)]^2$的极小值：

$$\begin{cases}\dfrac{\partial Q}{\partial \beta_0}=-2\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_i))=0\\[2ex]\dfrac{\partial Q}{\partial \beta_1}=-2\sum_{i=1}^{n}[y_i-(\beta_0+\beta_1 x_i)]x_i=0\end{cases}$$

$$\Rightarrow\quad \begin{cases}n\beta_0+n\bar{x}\beta_1=n\bar{y}\\ n\bar{x}\beta_0+(\sum_{i=1}^{n}x_i^{\,2})\beta_1=\sum_{i=1}^{n}x_i y_i\end{cases}$$

正规方程组

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

参数的最小二乘估计

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

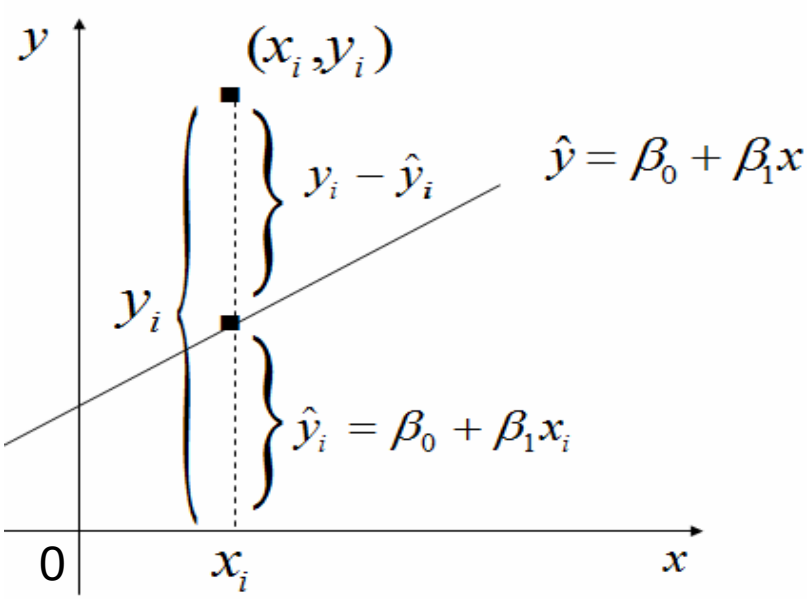
$$L_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

思考题:

- 求参数的最小二乘估计是否用到了线性回归的三个前提条件?

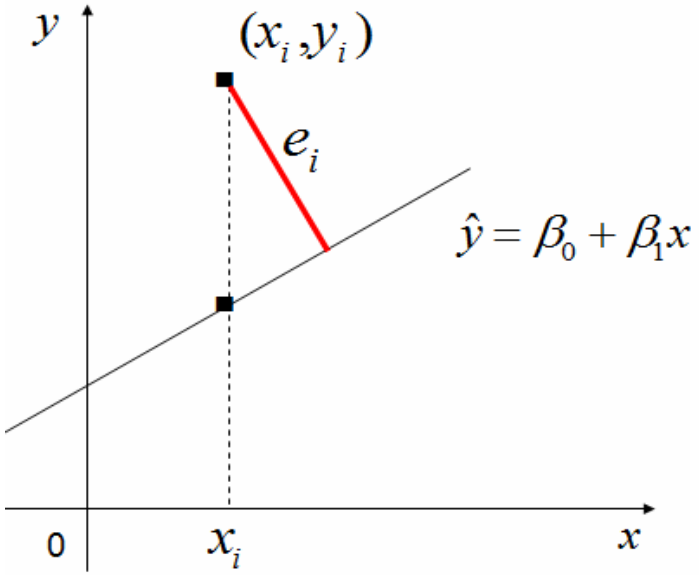
- 求一条与所有观测值的散点"最接近"（即Q极小）的直线.

  能否用其他的方法来刻画 "最接近"？



$$Q_{\min} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \Rightarrow \hat{\beta}_0, \hat{\beta}_1$$

经典回归



$$\tilde{Q}_{\min} = \sum_{i=1}^{n} e_i^2 \Rightarrow \tilde{\beta}_0, \tilde{\beta}_1$$

距离回归

回归分析中参数的极大似然估计

由 $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ （$\varepsilon_i \sim N(0,\sigma^2)$，易知：

$$y_i \sim N(\beta_0 + \beta_1 x_i,\ \sigma^2)$$

于是，（y_1，y_2，…，y_n）的联合密度函数为：

$$p(y_1,\ y_2,\ \ldots,y_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2}$$

取似然函数 $L(\beta_0,\ \beta_1,\ \sigma^2) = \dfrac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2}$

由 $\begin{cases} \dfrac{\partial \ln L}{\partial \beta_0} = 0 \\ \dfrac{\partial \ln L}{\partial \beta_1} = 0 \\ \dfrac{\partial \ln L}{\partial \sigma^2} = 0 \end{cases}$ $\Rightarrow$ $\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \dfrac{L_{xy}}{L_{xx}} \\ \hat{\sigma}^2 = \dfrac{1}{n}\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)^2 \end{cases}$

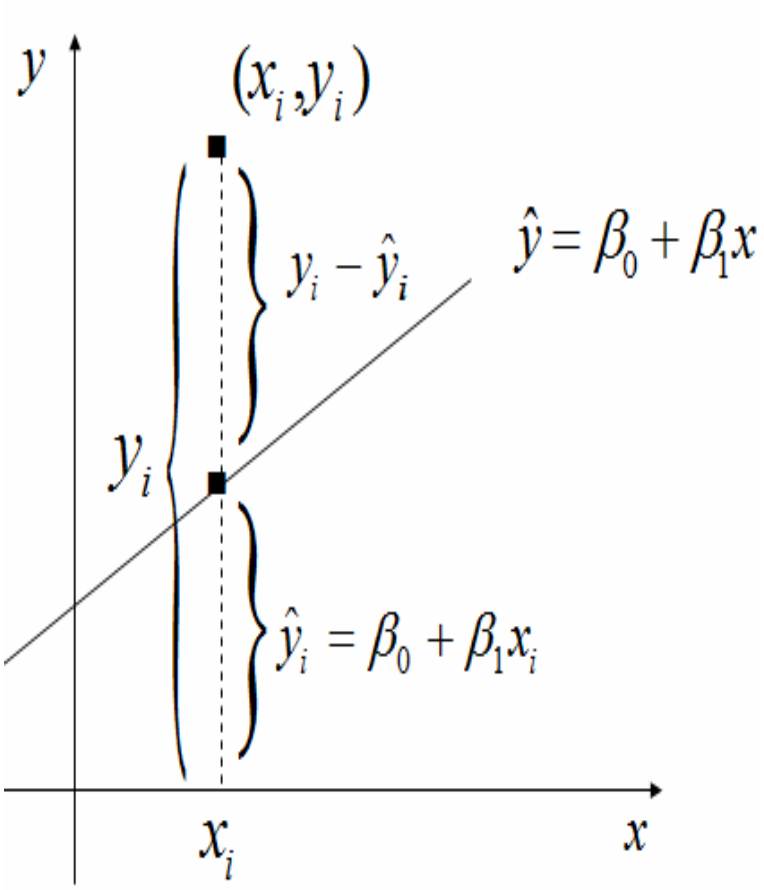
参数的极大似然估计与最小二乘估计相同

思考题：

可否用矩法估计的方法求解参数?

为什么?

参数的最小二乘估计(极大似然估计)效果如何?



y_i 的平均值 $\bar{y}$ 就等于
回归直线上对应的$\hat{y}_i$的平均值

证明: $\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i)$

$$= \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$= (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 \bar{x} = \bar{y}$$

残差平方和：

如果我们把求出的参数$\hat{\beta}_0$ ，$\hat{\beta}_1$代入 Q，得：

$$Q_{\min}=\sum_{i=1}^{n}[y_i-(\hat{\beta}_0+\hat{\beta}_1x_i)]^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \triangleq SS_e$$ ——称为残差平方和

注： SS_e越小，表示观测值点与回归直线越接近

当SS_e=0时，表示所有观测值点都在回归直线上

$$SS_e=L_{yy}-\hat{\beta}_1L_{xy}=L_{yy}-\hat{\beta}_1^2L_{xx}$$

证: $$SS_e=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}[y_i-(\hat{\beta}_0+\hat{\beta}_1x_i)]^2$$

$$=\sum_{i=1}^{n}[y_i-(\bar{y}-\hat{\beta}_1\bar{x})-\hat{\beta}_1x_i)]^2$$

$$=\sum_{i=1}^{n}[(y_i-\bar{y})-\hat{\beta}_1(x_i-\bar{x})]^2=L_{yy}-2\hat{\beta}_1L_{xy}+\hat{\beta}_1^2L_{xx}$$

$$=L_{yy}-\hat{\beta}_1L_{xy}=L_{yy}-\hat{\beta}_1^2L_{xx} \qquad (\because\ \hat{\beta}_1=L_{xy}/L_{xx})$$

估计标准差：

注意到我们已经证明：误差项$\varepsilon_i \sim N(0,\sigma^2)$ 中方差σ^2 的极大似然估计为

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \frac{SS_e}{n}$$

但这个估计不是无偏的，后面将证明σ^2 的无偏估计为$\frac{SS_e}{n-2}$

因此称 $\hat{\sigma} = \sqrt{\frac{SS_e}{n-2}}$ 为一元回归的估计标准差

注：估计标准差$\hat{\sigma}$ 越小回归效果越好

变元X与Y的样本相关系数：

$$r = \frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

对比 随机变量X与Y的相关系数

$$\rho=\frac{\mathrm{cov}(X,Y)}{\sqrt{DX\cdot DY}}=\frac{E(X-EX)(Y-EY)}{\sqrt{DX\cdot DY}}$$

- r是随机变量X与Y的相关系数 ρ 的矩法估计
- r是当$(X,Y)\sim N(\mu_1,\sigma_1^2;\mu_2,\sigma_2^2;\rho)$时, 参数$\rho$ 的极大似然估计

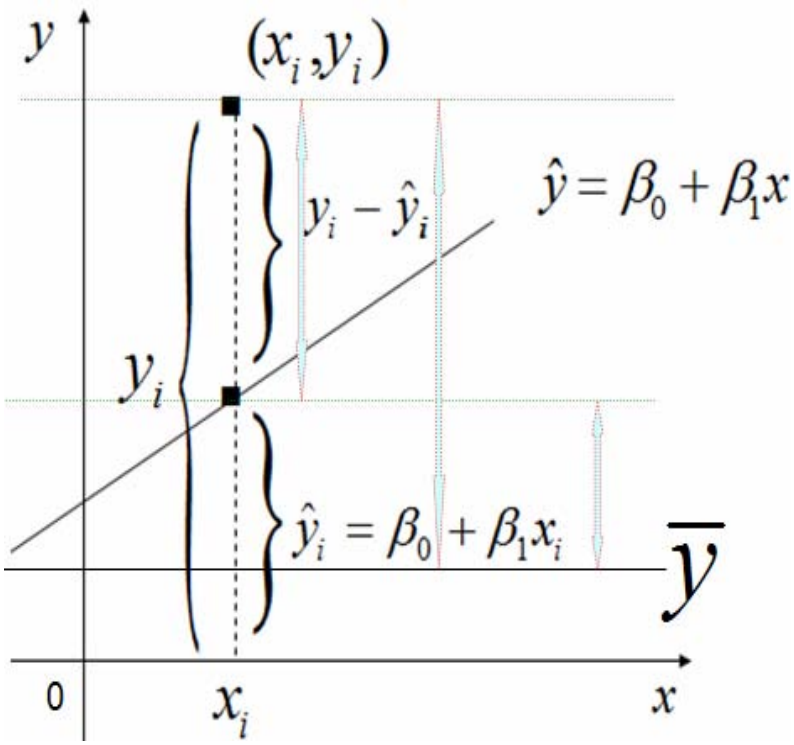
变元X与Y的相关系数r的性质：

1) $-1 \le r \le 1$

2) |r|越大，表示变元X与Y线性关系越强，反之，则表示线性关系越弱

3) r>0表示变元X与Y是正统计相关关系，即X越大则大体上Y也越大

r<0表示变元X与Y是负统计相关关系，即X越大而大体上Y会越小



如果记：$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$ ---总离差平方和

$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ ---回归平方和

及前面讲到的：

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

则可以证明：

$SS_T = SS_R + SS_e$---离差分解公式

利差分解公式：总离差平方和=回归平方和+残差平方和

证明：

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

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

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

$$= SS_e + 0 + SS_R$$

注：其中$2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$ 是根据正规方程导出的。

判定系数：

我们称回归平方和与总离差平方和的比值 $\frac{SS_R}{SS_T}$ 为可决系数

或判定系数(coefficient of determination)，记为：$r^2 = \frac{SS_R}{SS_T}$

判定系数 $\frac{SS_R}{SS_T}$ 记为 r^2, 与变元X与Y的样本相关系数 r 的平方，不就容易混淆了吗?

事实上: 样本相关系数的平方

$$r^2 = \left(\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}\right)^2 = \frac{L_{xy}^2}{L_{xx}L_{yy}} = \frac{L_{xy}}{L_{xx}}\frac{L_{xy}}{L_{yy}} = \hat{\beta}_1\frac{L_{xy}}{L_{yy}}$$

$$= \frac{L_{yy} - L_{yy} + \hat{\beta}_1 L_{xy}}{SS_T} = \frac{L_{yy} - SS_e}{SS_T} = \frac{SS_R}{SS_T}$$

注：一元回归的离差分解公式及可决系数的定义可直接推广到多元线性回归

说明：

$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ 反映了回归中自变量变差的贡献

$SS_e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ 反映了其它因素的影响

$r^2 = \mathbf{SS}_R / \mathbf{SS}_T$ 反映了回归方程对观测数据的拟和程度

例1 测量上海市1～3岁男孩的平均体重，得到数据如下：

年龄x_i（岁）	1.0	1.5	2.0	2.5	3.0
体重y_i（kg）	9.75	10.81	12.07	12.88	13.74

设 $y_i=\beta_0+\beta_1 x_i+\varepsilon_i$ ，$\varepsilon_i\sim N(0,\sigma^2)$ ，$i=1,2,\cdots,5$ ，$\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_5$ 相互独立。

求：（1） β_0,β_1 的最小二乘估计 $\hat{\beta}_0,\hat{\beta}_1$ ；

（2） 残差平方和 SS_e ，估计的标准差 $\hat{\sigma}$ ，样本相关系数 r 。

解：（1） $n=5$ ，$\bar{x}=2$ ，$L_{xx}=2.5$ ，$\bar{y}=11.85$ ，$L_{yy}=10.173$

$$L_{xy}=\sum_{i=1}^{n}x_i y_i-n\,\bar{x}\,\bar{y}=123.525-5\times 2\times 11.85=5.025$$

所以，回归方程为$\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x=7.83+2.01x$

（2） $SS_e=L_{yy}-\hat{\beta}_1 L_{xy}=10.173-2.01\times 5.025=0.07275$

$$\hat{\sigma}=\sqrt{\frac{SS_e}{n-2}}=\sqrt{\frac{0.07275}{5-2}}=0.1557 \qquad r=\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}=\frac{5.025}{\sqrt{2.5\times 10.173}}=0.9964$$

关于上述例1，请大家思考如下问题：

- 我们得到的回归方程有什么用？
- 根据哪些指标可以判断回归的效果？上述回归的效果如何？
- 上例中：年龄为自变量（控制变量），体重为因变量（响应变量），回归方程为： $y = 7.83+2.01x$ ，那么据此方程得： $x = (y -7.83)/2.01$ ，它可否视为把体重作为自变量，年龄作为因变量的回归方程？
- 对于任意给定的一组数值（x_i, y_i） $i=1,2,\ldots,n$，比如 x_i 表示第i天的最高气温， y_i表示第i天股市的收盘指数，是否都可以像例1一样代入参数的公式并求出回归方程？
- 如果观测值较多，直接手算比较复杂，如何借助计算机求解回归方程？

关于问题1：回归方程有什么用途？

回归方程的主要用途是预测和控制，比如根据上例的回归方程

y = 7.83+2.01x ，我们可以预测 x=2.2(岁）时儿童的体重为:

y = 7.83+2.01*2.2=12.252(kg)-----这是y的点估计，我们还可以得到y的区间估计。

对于一元线性回归模型 $y=\beta_0+\beta_1 x+\varepsilon$ ，其中误差项满足正态性，独立性，及方差齐性的条件， 给定 x_0，则对应 y_0 的点估计为 $\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1 x_0$ ；当 n 充分大时，y_0 置信水平为$1-\alpha$ 的置信区间可近似表示为 $[\hat{y}_0-\hat{\sigma}u_{1-\frac{\alpha}{2}},\ \hat{y}_0+\hat{\sigma}u_{1-\frac{\alpha}{2}}]$

此外，我们还可以求出参数 β_0和β_1的区间估计

β_0和β_1置信水平为$1-\alpha$的置信区间分别为：

$$[\hat{\beta}_0-\hat{\sigma}t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{L_{xx}}},\ \hat{\beta}_0+\hat{\sigma}t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{L_{xx}}}]$$

和 $$[\hat{\beta}_1-\hat{\sigma}t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\frac{1}{L_{xx}}},\ \hat{\beta}_1+\hat{\sigma}t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\frac{1}{L_{xx}}}]$$

关于问题2：哪些指标可以判断回归的效果？

如下指标都可以直接或间接用来表示回归的效果：

参差平方和 SSe

估计标准差 $\hat{\sigma}$

相关系数 r

判定系数 r^2

修正判定系数 $R_a^2 = 1-(1-r^2)\frac{n-1}{n-p-1}$

其中p为自变元个数

从例1第二问的结果看，该例回归的效果还是很好的

关于问题3：能否由体重关于年龄的回归方程：

$$y = 7.83+2.01x$$ ， 得出年龄关于体重的回归方程：

$$x = (y - 7.83)/2.01 = 0.4975y - 3.8955$$ ？

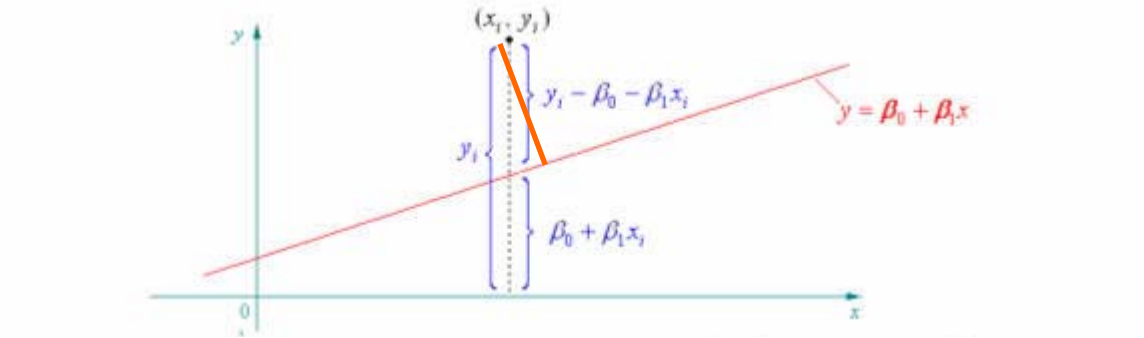
不可以。事实上，如果把体重作为自变量年龄作为因变量，

代入一元回归的公式，得：$x = 0.4939y - 3.853$;

二者为何不同呢？

因为经典回归中，自变量与响应变量的地位是不等同的

而对距离回归，即通过各散点到回归函数的距离平方和最小来求出回归参数，此时自变量与响应变量的地位是等同的，这种情况下是可以直接从 y关于 x 的回归方程解出 x 关于 y 的回归方程的



关于问题4： 对于任意给定的一组数值（x_i， y_i） i=1,2,...,n，是否都可以求变量的回归方程？

可以代入参数最小二乘估计的公式求出变元的回归方程，但是，如果变元 X 和 Y 没有统计相关关系，这样求出的回归方程是没有意义的（如气温与股票点数）；而如果回归模型的三个条件，即正态性，独立性，方差齐性 不满足，我们就无法对参数的概率特性（分布，区间估计 等）作出判断。

直观地说，如果根据变元 X 和 Y 的观测值算出的相关系数的绝对值越大（越接近1），即表示变元 X 和 Y 线性关系越强，这时拟合观测值（x_i, y_i)的回归方程越有意义

那么，相关系数的绝对值要达到多大才可以求回归方程呢？

在统计上，我们是用假设检验的方法来判定变元的线性关系是否显著，因为检验的统计量服从F分布，因此这个检验叫F检验

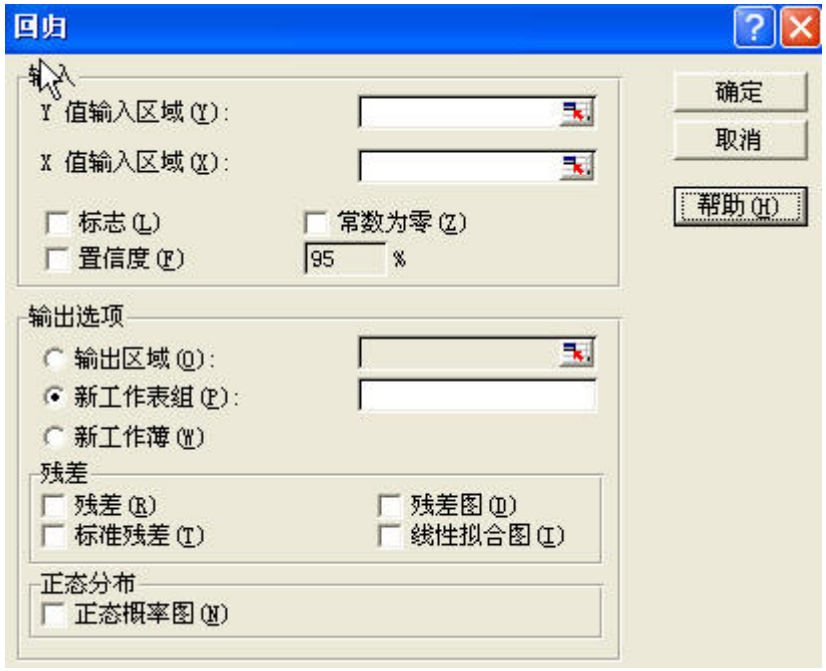
关于问题5： 如何借助计算机算法进行回归分析？

各种统计软件都有回归分析的功能，比如SAS，SPSS，R，包括MATLAB的统计包 等，这里我们介绍EXCEL的回归分析功能

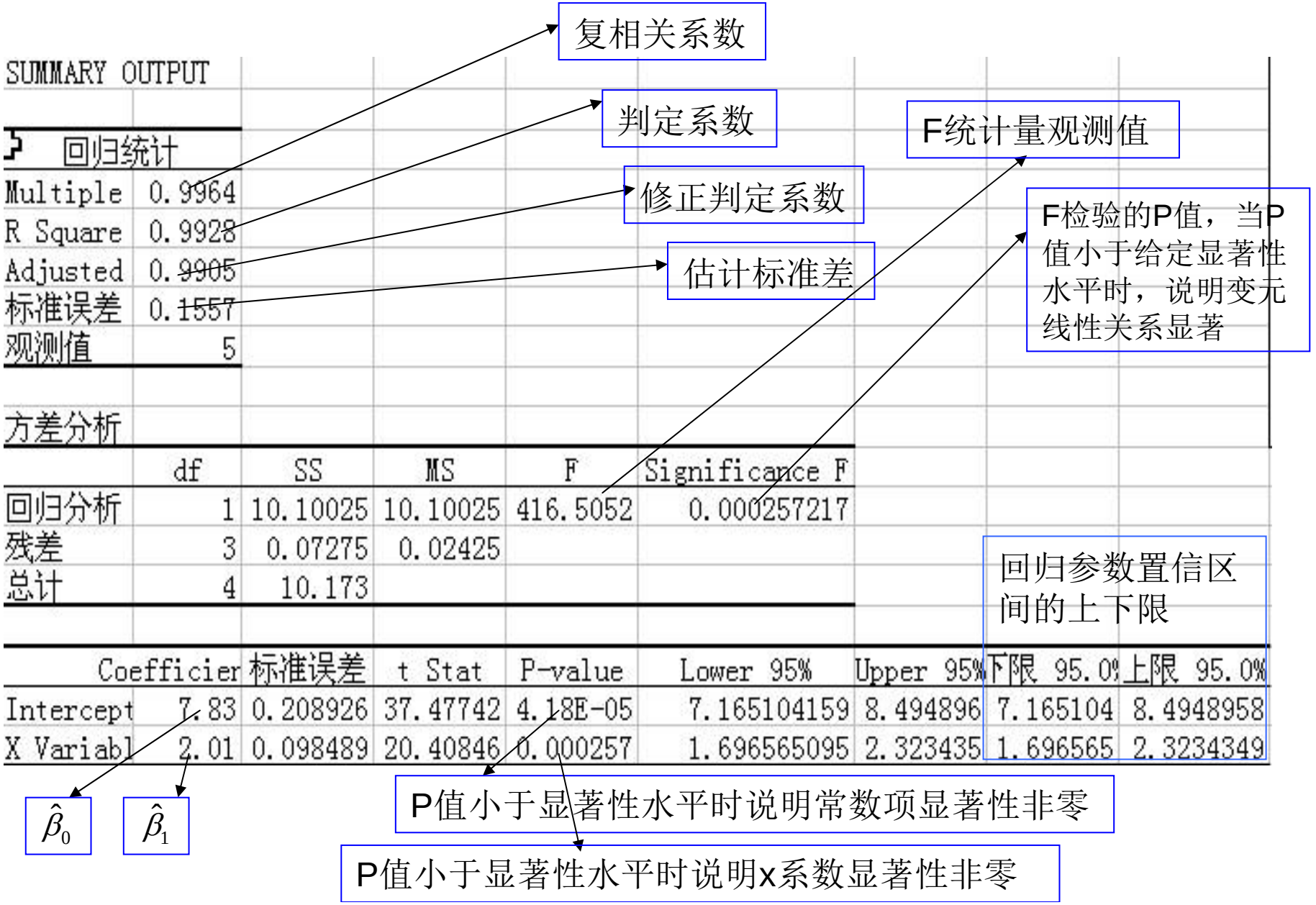
操作步骤（多元回归同样操作，但利用EXCEL多元回归分析时自变元个数不能超过16个）：

1）把数据输入EXCEL表

2）点工具菜单 → 加载宏 → 数据分析 → 回归



对例1中数据的EXCEL回归分析结果：



SUMMARY OUTPUT

回归统计	
Multiple	0.9964
R Square	0.9928
Adjusted	0.9905
标准误差	0.1557
观测值	5

方差分析

	df	SS	MS	F	Significance F
回归分析	1	10.10025	10.10025	416.5052	0.000257217
残差	3	0.07275	0.02425		
总计	4	10.173			

	Coefficier	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	7.83	0.208926	37.47742	4.18E-05	7.165104159	8.494896	7.165104	8.4948958
X Variabl	2.01	0.098489	20.40846	0.000257	1.696565095	2.323435	1.696565	2.3234349

例2：恩格尔系数（食品支出与收入之比）的估算

已知人均月收入X与人均食品月支出Y的15组抽样数据如下，求恩格尔系数：

X	1020	960	970	1020	910	1580	540	830	1230	1060	1290	1380	810	920	640
Y	270	260	250	280	270	360	190	260	310	310	340	380	270	280	200

分析：根据给定数据，先找出X，Y的回归函数，再根据回归函数来估计恩格尔系数

解：利用EXCEL进行回归分析，得：

1020	270	SUMMARY OUTPUT							
960	260								
970	250	回归统计							
1020	280	Multiple	0.94145						
910	270	R Square	0.886328						
1580	360	Adjusted	0.877584						
540	190	标准误差	18.28581						
830	260	观测值	15						
1230	310								
1060	310	方差分析							
1290	340		df	SS	MS	F	Significance F		
1380	380	回归分析	1	33893.18	33893.18	101.364	1.66314E-07		
810	270	残差	13	4346.82	334.3707				
920	280	总计	14	38240					
640	200								
			Coefficien	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
		Intercept	99.87161	18.69586	5.341911	0.00013	59.48167559	140.2615	59.48168	140.2615
		X Variabl	0.180206	0.017899	10.06797	1.7E-07	0.141537857	0.218875	0.141538	0.218875

于是，得X，Y的回归方程为 $\hat{Y}=99.8716+0.1802X$

即： $\frac{\hat{Y}}{X}=\frac{99.8716}{X}+0.1802$

即恩格尔系数约为0.1802,且恩格尔系数会随收入的增大而变小