概率论与数理统计

第八章

应用回归分析

回归分析的研究对象

现实世界中变量之间的关系并不总是可以用函数关系(*<u>自变量确定,因变量唯一</u>*)来表示的

比如:

1) 家庭收入与家庭支出的关系
2) 父母身高与子/女身高的关系
3) 平时作业成绩与最后的考试成绩的关系
4) 银行利率与股票指数的关系

- **统计相关关系:**

经验和统计数据表明某些变量的取值相互之间是有关系的,不是完全无关的,这种关系称为**统计相关关系**

- **回归分析及回归方程:**

回归分析就是研究变量间的统计相关关系一种统计方法.

根据变元的统计数据,用一个函数来近似变元间的统计相关关系,这个函数叫**回归方程**或**回归函数**

1886 年，高尔顿发表论文《遗传中向平均身高回归的现象》。高尔顿与皮尔逊合作，一起研究这个课题。他们收集了 1078 对父亲和儿子身高的数据：

父身高　子身高

(x_i, y_i)，$i = 1, 2, \cdots, 1078$

得到直线的方程为

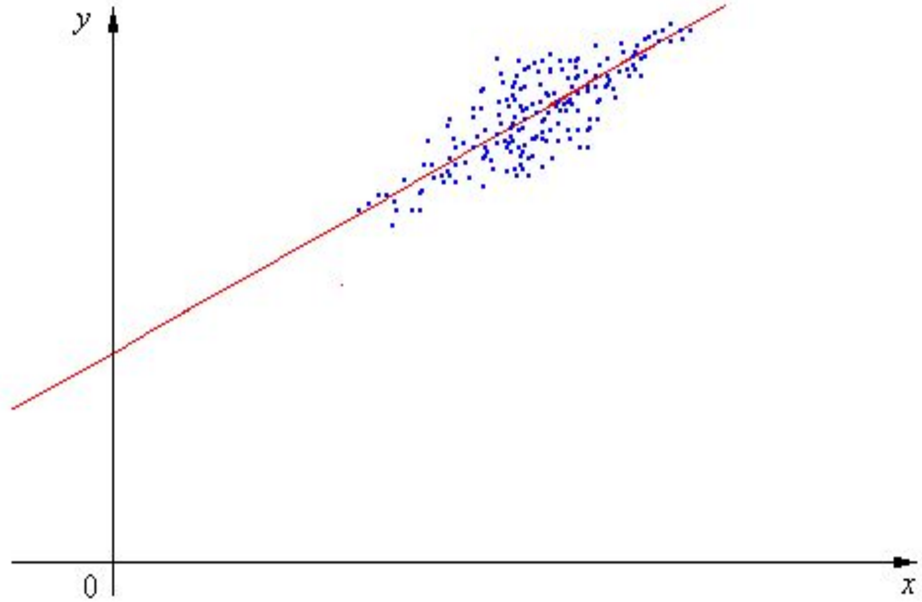
$\hat{y} = 0.8567 + 0.516x$　（单位：米）

例：$x = 1.900 \to \hat{y} = 1.837$

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

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

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



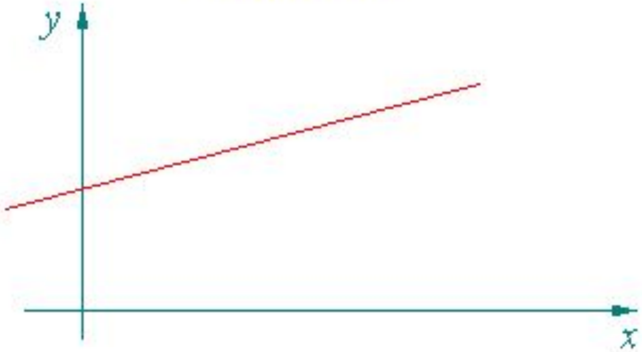
本例中, 父亲身高与儿子身高的关系就是*统计相关关系*

上述高尔顿得到的近似直线方程就是*回归方程*

回归方程，可以是线性的，也可以是非线性的，当回归方程为线性时，称为**线性回归**（Linear Regression），当回归方程为非线性时，称为**非线性回归**（Nonlinear Regression）。

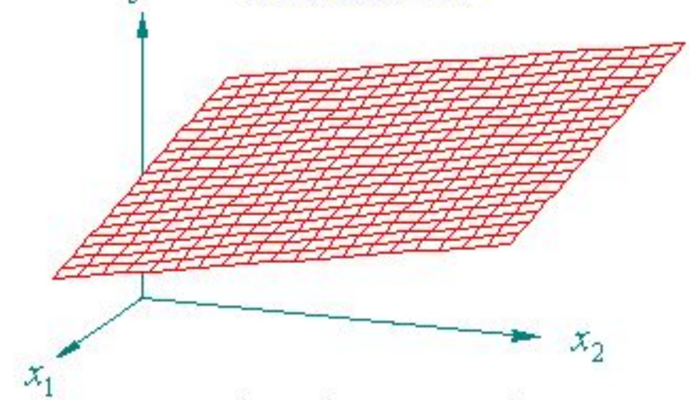
在回归方程中，可以只有一个自变量，也可以有多个自变量，只有一个自变量的回归称为**一元回归**（Simple Regression），有多个自变量的回归称为**多元回归**（Multiple Regression）。

一元线性回归



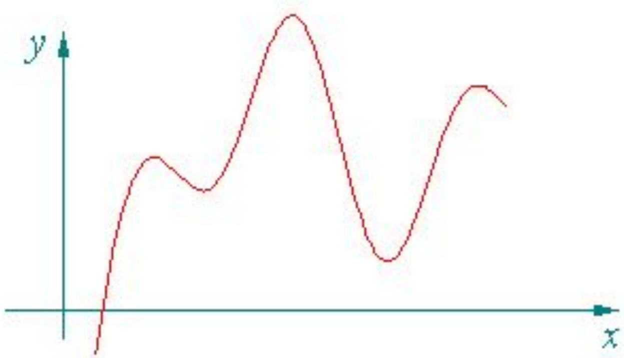
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

多元线性回归



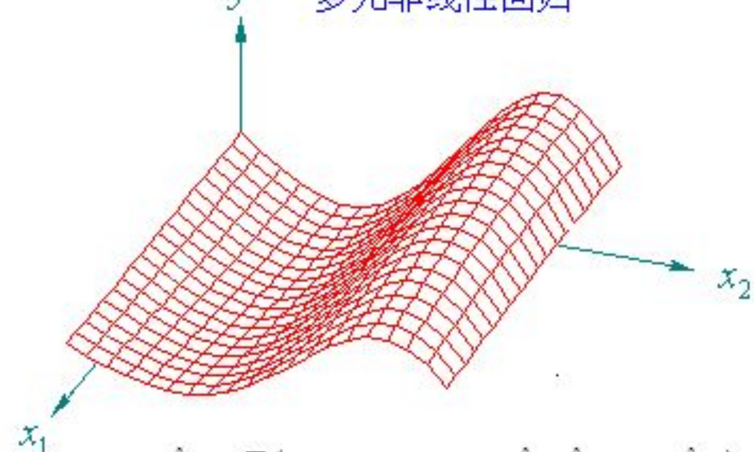
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_m x_m$$

一元非线性回归



$$\hat{y} = F(x; \hat{a}_1, \hat{a}_2, \cdots, \hat{a}_p)$$

多元非线性回归



$$\hat{y} = F(x_1, x_2, \cdots, x_m; \hat{a}_1, \hat{a}_2, \cdots, \hat{a}_p)$$

§8.1　一元线性回归

- 一元线性回归的模型:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

其中,X为确定性变量,它是可以测量和控制的,也称解释变量或自变量;

Y为被解释变量或响应变量

β_0和β_1为未知的待估计参数

ε为误差项,它表示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, . . . , n）代入上述一元线性回归模型，得：

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

为了从上述这组等式中解出未知参数β_0和β_1，及判断他们具有哪些性质，通常我们要求随机项ε_i满足下述三个前提条件：

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

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

这三个性质是我们回归分析的前提，一般说来这三个性质是满足或近似满足的，比如正态性，我们知道误差的分布一般是服从正态分布的（事实上正态分布就是高斯研究误差时提出的）。独立性和方差齐性是为了便于分析的附加条件，严格说来，在讨论实际问题时，我们还需要对这三个条件进行检验和验证：

1) 正态性检验方法：本书7.3.2节分布的检验，或正态分布概率纸检验

2) 独立性检验方法：独立性χ^2检验，本书8.3节参差分析

3）方差齐性检验：本书7.2.2节讲了两个随机变量等方差的检验，

多个随机变量等方差的检验见本书8.3节参差分析

回到我们的一元线性回归模型：

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

其中误差项满足：

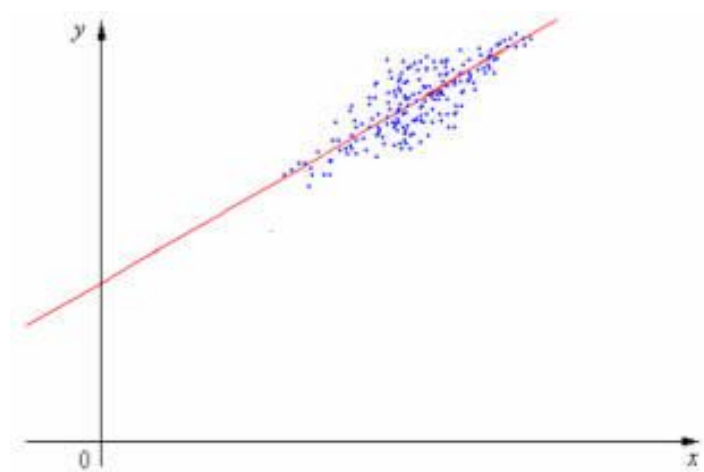
1) 正态性： $\varepsilon_i \sim N(0, \sigma^2)$

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

观测值（x_i, y_i）即散点图中的各个点，如果没有随机误差项ε_i，这些点都将落在直线（回归方程）上，因为ε_i的不同取值，才导致了y_i可能偏离了回归直线。因为ε_i是随机变量，因此

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ 也都是随机变量



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：

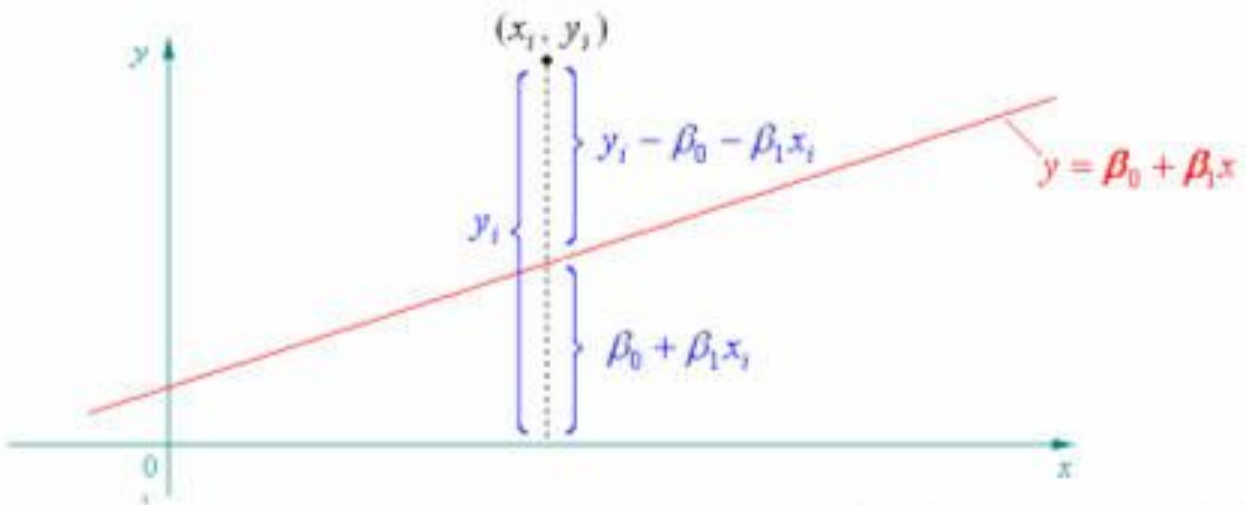
$$p(y_1, y_2, ..., y_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2}$$

其中，$\beta_0, \beta_1, \sigma^2$均为未知参数。根据极大似然估计的方法：

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

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

我们导出了参数的极大似然估计，但是，历史上高尔顿是用我们高等数学中所学过的最小二乘法导出的，因此，一般称之为最小二乘估计



问题 已知 (x_i, y_i)，$i=1,2,\cdots,n$，求常数 β_0, β_1 的估计 $\hat{\beta}_0, \hat{\beta}_1$，使得当 $\beta_0 = \hat{\beta}_0$，$\beta_1 = \hat{\beta}_1$ 时，$Q=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2$ 达到最小。

$$推导：\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = 0 \\[2ex] \dfrac{\partial Q}{\partial \beta_1} = 0 \end{cases} \quad \Rightarrow \quad \begin{cases} \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \\[2ex] \hat{\beta}_1 = \dfrac{L_{xy}}{L_{xx}} \end{cases}$$

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

$$Q_{\min} = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \triangleq SSE$$ ——称为残差平方和

显然，SSE越小，表示观测值距回归直线越近，特别地：

当 SSE=0 时，表示所有观测值的点都在回归直线上。

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

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

但这个估计不是无偏的，可以证明σ^2的无偏估计为$\dfrac{SSE}{n-2}$，

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

估计标准差越小，即SSE越小，它也表示回归效果越好

除残差平方和SSE，估计标准差$\hat{\sigma}$可以表示回归效果外，

我们还可以用相关系数来表示回归的效果

变元X与Y的相关系数的定义是： $$R=\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2\sum_{i=1}^{n}(y_i-\overline{y})^2}}$$

对比我们曾学过的随机变量X与Y的相关系数

$$r=\frac{\operatorname{cov}(X,Y)}{\sqrt{DX\bullet DY}}=\frac{E(X-EX)(Y-EY)}{\sqrt{DX\bullet DY}}$$

会发现他们形式上很象。事实上，变元X与Y的相关系数r的定义就是把（X， Y）视为服从二维正态分布时，其相关系数ρ的极大似然估计

变元X与Y的相关系数 $R=\dfrac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}$ 与随机变量的相关系数也有类似的性质：

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：

$$SST \triangleq \sum_{i=1}^{n}(y_i - \bar{y})^2 \text{---总离差平方和}$$

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

及前面讲到的：

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

则可以证明：

$$SST = SSR + SSE \text{---离差分解公式}$$

证明： $SST \triangleq \sum_{i=1}^{n}(y_i-\bar{y})^2$

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

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

$$=SSE+0+SSR$$

注其中： $\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})=0$ 是根据

所谓的正规方程，即：

$$\begin{cases}\dfrac{\partial Q}{\partial \beta_0}=0\\[2ex]\dfrac{\partial Q}{\partial \beta_1}=0\end{cases}$$

导出的。

离差分解公式　　ＳＳＴ＝ＳＳＲ＋ＳＳＥ

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

或判定系数(coefficient of determination)，记为：$R^2 \triangleq \frac{SSR}{SST}$

注:

1） 可以证明可决系数$\frac{SSR}{SST}$一定等于变元X与Y相关系数R的平方，

因此，可记 $R^2 \triangleq \frac{SSR}{SST}$ （证明略，提示利用正规方程）

2）离差分解公式中，SSR=$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$表示回归方程 $\hat{y}_i$的离差平方和（$\hat{y}_i$的均值等于$\bar{y}$），

SSE是由随机误差造成的，ε_i的方差σ^2越大则SSE会越大，（$\frac{SSE}{n-2}$是σ^2的无偏估计）

3）上述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

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 $n = 5$ ，$\bar{x} = 2$ ，$L_{xx} = 2.5$ ，$\bar{y} = 11.85$ ，$L_{yy} = 10.173$ ，$L_{xy} = 123.525 - 5 \times 2 \times 11.85 = 5.025$ 。

（1） $$\hat{\beta}_1 = \frac{L_{xy}}{L_{xx}} = \frac{5.025}{2.5} = 2.01 ,\qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 11.85 - 2.01 \times 2 = 7.83 。$$

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

（2） $$SS_e = L_{yy} - \hat{\beta}_1 L_{xy} = 10.173 - 2.01 \times 5.025 = 0.07275 ,\quad \hat{\sigma} = \sqrt{\frac{SS_e}{n-2}} = \sqrt{\frac{0.07275}{5-2}} = 0.1557 ,$$

$$r = \frac{L_{xy}}{\sqrt{L_{xx} L_{yy}}} = \frac{5.025}{\sqrt{2.5 \times 10.173}} = 0.9964 。$$

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：

残差平方和 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：

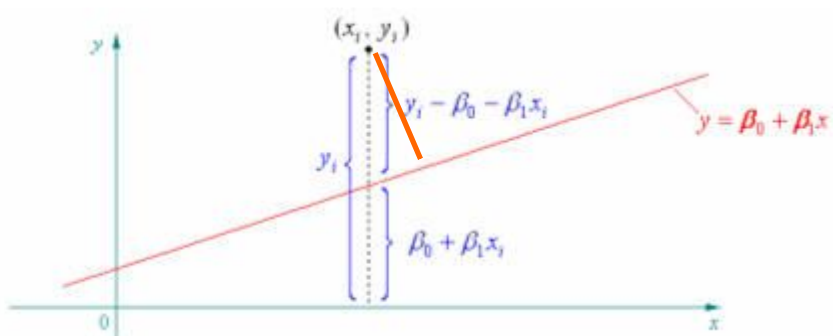
$x = (y -7.83)/2.01=0.4975y – 3.8955$ ？

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？

因为我们这里介绍的一元回归模型中，自变量与响应变量的地位是不等同的

还有一种回归，叫距离回归，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）$i=1, 2, \ldots, n$，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（证明略），因此这个检验叫F检验

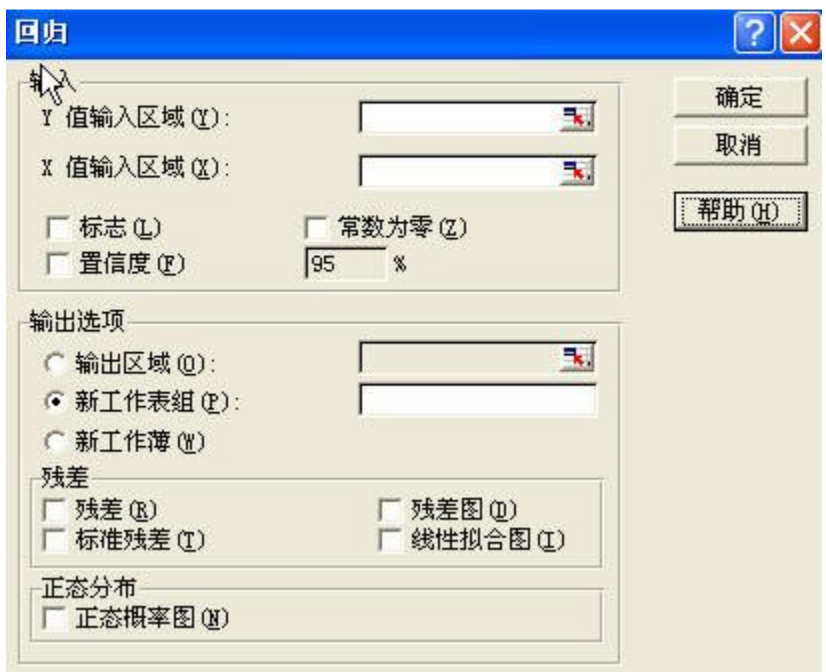
关于问题4：如何借助计算机算法进行回归分析？

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

相关系数

判定系数

修正判定系数

估计标准差

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			

F统计量观测值

F检验的P值，当P值小于给定显著性水平时，说明变元线性关系显著

	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

回归参数置信区间的上下限

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

P值小于显著性水平时说明常数项显著性非零

P值小于显著性水平时说明x系数显著性非零

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	ignificance 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

§ 8.2　多元线性回归

设自变量 $x_1, x_2, \cdots, x_m$ 与因变量 y 之间，有下列关系：

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon \quad ,$$

其中，$\beta_0, \beta_1, \cdots, \beta_m$ 是常数，$\varepsilon \sim N(0, \sigma^2)$ 是表示误差的随机变量。

对 $x_1, x_2, \cdots, x_m$， y 进行 n 次观测，得到一组观测值：

$$(x_{i1}, x_{i2}, \cdots, x_{im}, y_i) \ , \ i = 1, 2, \cdots, n \ 。$$

即有方程组

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \cdots + \beta_m x_{1m} + \varepsilon_1 \\ \qquad \vdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \cdots + \beta_m x_{nm} + \varepsilon_n \end{cases} ,$$

其中 $\varepsilon_i \sim N(0, \sigma^2)$ ，$i = 1, 2, \cdots, n$ ，$\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ 相互独立。

多元线性回归，就是要求出未知常数 $\beta_0,\beta_1,\cdots,\beta_m$ 的最小二乘估计 $\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_m$ ，使得回归方程 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_mx_m$ 能够尽可能精确地将自变量 $x_1,x_2,\cdots,x_m$ 与因变量 y 之间的统计相关关系表达出来。

为了简单起见，我们将它写成矩阵向量形式。

令 $Y=\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$， $X=\begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix}$， $\beta=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}$， $e=\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$ 。

则上述方程组可以简写成

$$Y=X\boldsymbol{\beta}+e，\ e\sim N_n(\mathbf{0},\sigma^2 I)，$$

其中，$N_n(\mathbf{0},\sigma^2 I)$ 表示 n 元正态分布，$\mathbf{0}=\begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$是数学期望向量，$\sigma^2 I=\begin{bmatrix} \sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2 \end{bmatrix}$是协方差矩阵。

显然这时有

$$Y=X\boldsymbol{\beta}+e\sim N_n(X\boldsymbol{\beta},\sigma^2 I)。$$

这就是多元线性回归的数学模型。

令：$$Q=\sum_{i=1}(y_i-\beta_0-\beta_1 x_{i1}-\cdots-\beta_m x_{im})^2$$

$$=\begin{bmatrix} y_1-\beta_0-\beta_1 x_{11}-\cdots-\beta_m x_{1m} \\ \vdots \\ y_n-\beta_0-\beta_1 x_{n1}-\cdots-\beta_m x_{nm} \end{bmatrix}^T \begin{bmatrix} y_1-\beta_0-\beta_1 x_{11}-\cdots-\beta_m x_{1m} \\ \vdots \\ y_n-\beta_0-\beta_1 x_{n1}-\cdots-\beta_m x_{nm} \end{bmatrix}$$

$$=(Y-X\beta)^T(Y-X\beta)=Y^TY-2\beta^TX^TY+\beta^TX^TX\beta$$ 。

可以通过求偏导数、解下列方程组的方法，来确定 Q 的最小值点：$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0}=0 \\ \dfrac{\partial Q}{\partial \beta_1}=0 \\ \cdots\cdots \\ \dfrac{\partial Q}{\partial \beta_m}=0 \end{cases}$$

解方程组得参数的 $\boxed{\text{最小二乘估计}\quad \hat{\beta}=(X^TX)^{-1}X^TY}$。 此外：

$\boxed{\text{残差平方和}\ SS_e=Y^TY-\hat{\beta}^TX^TY}=Y^TY-\hat{\beta}^TX^TX\hat{\beta}$，$\boxed{\text{估计的标准差}\ \hat{\sigma}=\sqrt{\dfrac{SS_e}{n-m-1}}}$，

$$\boxed{\text{多重相关系数}\ r=\sqrt{1-\frac{SS_e}{L_{yy}}}\ 。}$$

$\boxed{r^2=\dfrac{SSR}{SST}\text{——复判定系数}}$

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

例3：试根据表中居民月收入X_1（单位百元）及某商品的单价X_2（单位十元）来拟合该商品的需求量Y（单位百件）的函数

需求 Y	月收入 X_1	商品单价 X_2
10	5	2
10	7	3
15	8	2
13	9	5
14	9	4
20	10	3
18	10	4
24	12	3
19	13	5
23	15	4

解：利用EXCEL进行回归分析，得：

SUMMARY OUTPUT

回归统计	
Multiple R	0.937724
R Square	0.879326
Adjusted R S	0.844848
标准误差	1.966842
观测值	10

方差分析

	df	SS	MS	F	Significance F
回归分析	2	197.3207	98.66036	25.50373	0.00061045
残差	7	27.07928	3.868468		
总计	9	224.4			

	Coefficien	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	4.587509	2.51998	1.820455	0.111494	-1.371291467	10.54631	-1.37129	10.54631
X Variable 1	1.868468	0.26961	6.930263	0.000225	1.230942306	2.505994	1.230942	2.505994
X Variable 2	-1.79957	0.732946	-2.45526	0.043769	-3.532711935	-0.06643	-3.53271	-0.06643

即拟合的函数为 $\hat{Y} = 4.5875 + 1.8685X_1 - 1.7996X_2$

因 F 检验的 P 值为0.00061，小于显著性水平0.05，说明 Y 与 X_1 和 X_2 有显著的线性相关关系，即所求的回归方程在显著性水平0.05下是成立的，有意义的。

需要说明的是：

多元回归的参数 $\hat{\beta}=(X^TX)^{-1}X^TY$，

如果X^TX不可逆，就无法求出 $\hat{\beta}$

根据线性代数的知识我们知道：

X^TX不可逆等价于矩阵X的列向量组（对应各自变元观测值向量）线性相关，即自变元之间存在线性相关关系（有“多余”的自变元），这在统计上叫复共线性问题，解决的方法是踢除“多余”的自变元，踢除的方法叫逐步回归，这是可以通过统计软件直接实现的

逐步回归，多元回归参数的检验，预测，参差分析，及应用案例等 略