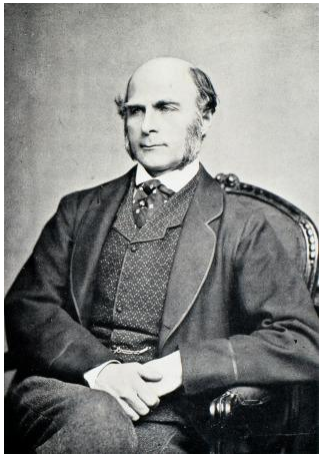
哈爾濱工業大學

第39讲 一元线性回归分析

什么是回归分析 (Regression)

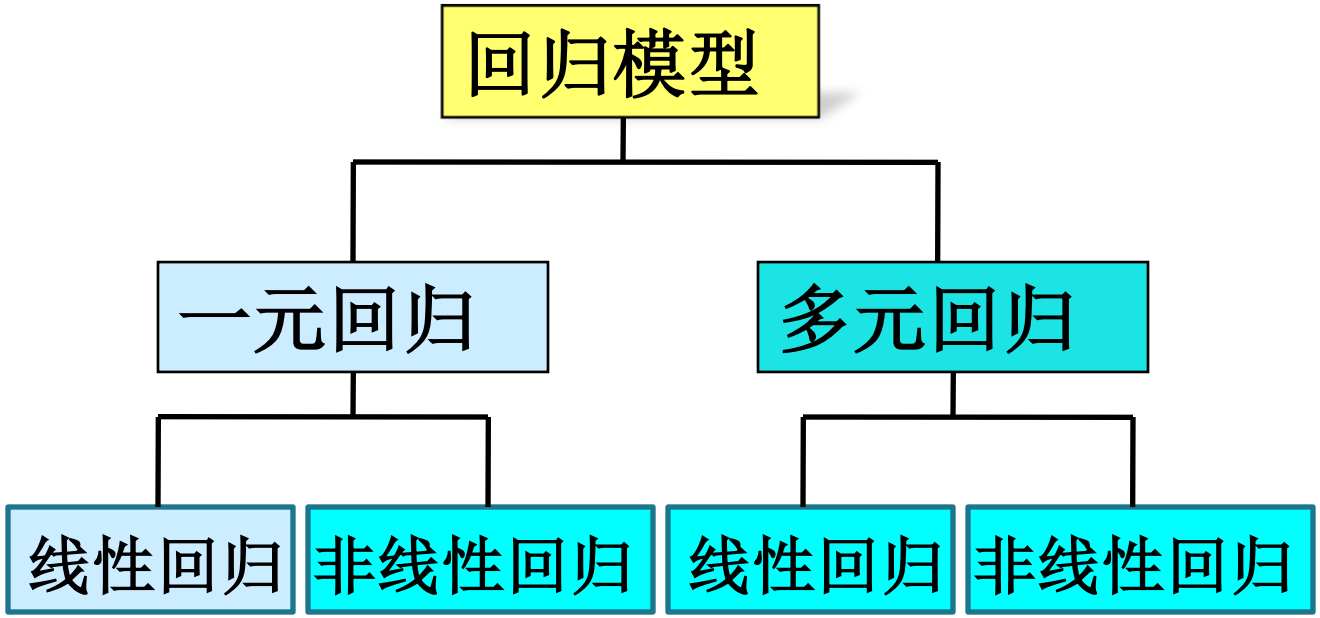
回归这个术语是由英国统计学家Galton在19世纪末,研究孩子身高与父母身高间的关系时提出的. Galton把这种孩子的身高向中间值靠近的趋势称之为一种回归效应，而他发展的研究两个数值变量的方法称为回归分析.



什么是回归分析 (Regression)

- 从一组样本数据出发，确定变量之间的数学关系式.
- 对数学关系式的可信程度进行各种统计检验，并从影响因变量的诸多自变量中找出哪些变量的影响显著，哪些不显著.
- 利用所求的关系式，根据一个或几个变量的取值来预测或控制另一个特定变量的取值，并给出这种预测或控制的精度.

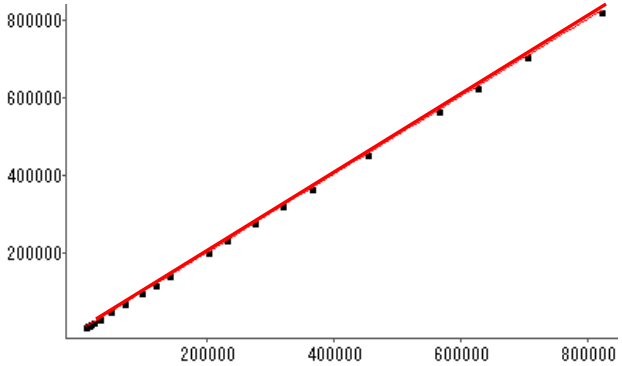
回归模型的类型



变量与变量间的两类关系

（1）确定性的—函数关系

当给定自变量一个值时，因变量有确定的值与之相对应. 例如，$y = x + 628.98$

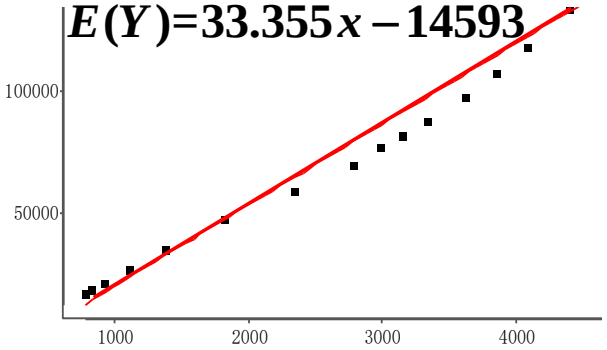


函数关系

变量与变量间的两类关系

（2）非确定性的—相关关系

当自变量取一定的值时，因变量取值具有随机性. 例如，小麦的单位面积施肥量x与单位面积产量Y之间的关系是不确定性的, 但有一定的趋势.



相关关系

一元线性回归过程

- 建模过程
  - 参数a, b的估计
  - 误差估计 σ^2
- 回归方程的显著性检验
  - F检验
  - t检验
  - r检验
- 回归方程的应用——预测与控制

一元线性回归模型

例1 根据《中国统计年鉴》2007年版，统计1978至2006年间用于科学研究的总费用与科研基建费数据如下：

8-7　国家财政用于科学研究的支出

单位：亿元

年 份	合 计	科研基建费
1978	52.89	6.66
1980	64.59	11.27
1985	102.59	18.83
1990	139.12	17.47
1991	160.69	18.40
1992	189.26	24.55
1993	225.61	33.95
1994	268.25	36.06
1995	302.36	38.00
1996	348.63	48.55
1997	408.86	42.74
1998	438.60	47.28
1999	543.85	52.89
2000	575.62	61.52
2001	703.26	63.37
2002	816.22	69.99
2003	975.54	111.06
2004	1095.34	95.90
2005	1334.91	112.50
2006	1688.50	134.40

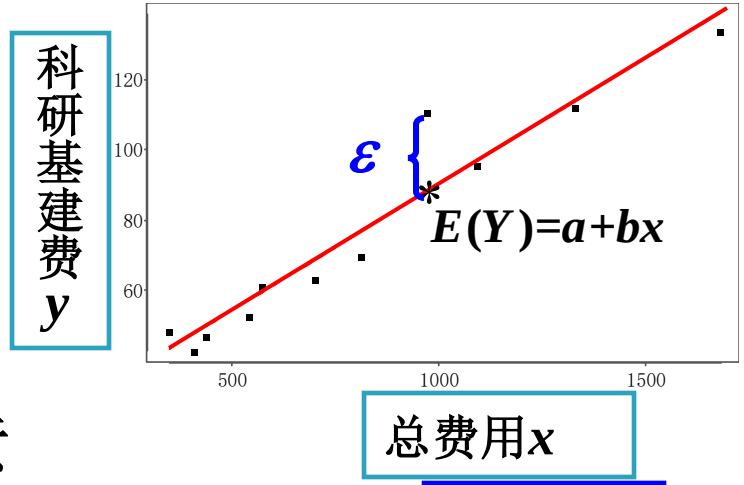
注：表中科研基建费由国家统计局测算，2006年数据为初步数

一元线性回归模型

由散点图可知：科研基建费由两部分组成：

$$Y = a + bx + \varepsilon \quad (1)$$

线性部分　随机因素



$\varepsilon \sim N(0,\sigma^2)$，$a,b,\sigma^2$为未知参数，

称(1)式为一元线性回归模型，a和b为回归系数.

称 $\hat{y} = \hat{a} + \hat{b}x$ 为回归方程.

一元线性回归模型的建立

用(x,Y)的 n 次独立观测值$(x_1,y_1),(x_2,y_2),\cdots$ (x_n,y_n)建立一元线性回归模型

$$\begin{cases} y_i = a + bx_i + \varepsilon_i, (i = 1,\cdots,n) \\ \varepsilon_i \sim N(0,\sigma^2), \text{相互独立} \\ a,b,\sigma^2 \text{为未知参数.} \end{cases}$$

要求回归方程$\hat{y} = \hat{a} + \hat{b}x$ ， 需估计参数a,b,σ^2.

参数 a,b 的最小二乘估计

设n组观测值 $(x_1,y_1),\cdots(x_n,y_n)$ 有

模型 $\begin{cases} y_i = a + bx_i + \varepsilon_i \\ \varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n \text{独立同分布} N(0,\sigma^2) \end{cases}$

使$\sum_{i=1}^{n}\varepsilon_i^2$最小的$\hat{a}$，$\hat{b}$为$a$，$b$的最小二乘估计.

$$\text{令} L = L(a,b) = \sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}(y_i - a - bx_i)^2$$

$$\begin{cases} \dfrac{\partial L}{\partial a} = -2\sum_{i=1}^{n}(y_i - a - bx_i) \\ \dfrac{\partial L}{\partial b} = -2\sum_{i=1}^{n} x_i(y_i - a - bx_i) \end{cases}$$

正规方程：

$$\begin{cases} n\hat{a} + \hat{b}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \hat{a}\sum_{i=1}^{n} x_i + \hat{b}\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

最小二乘估计

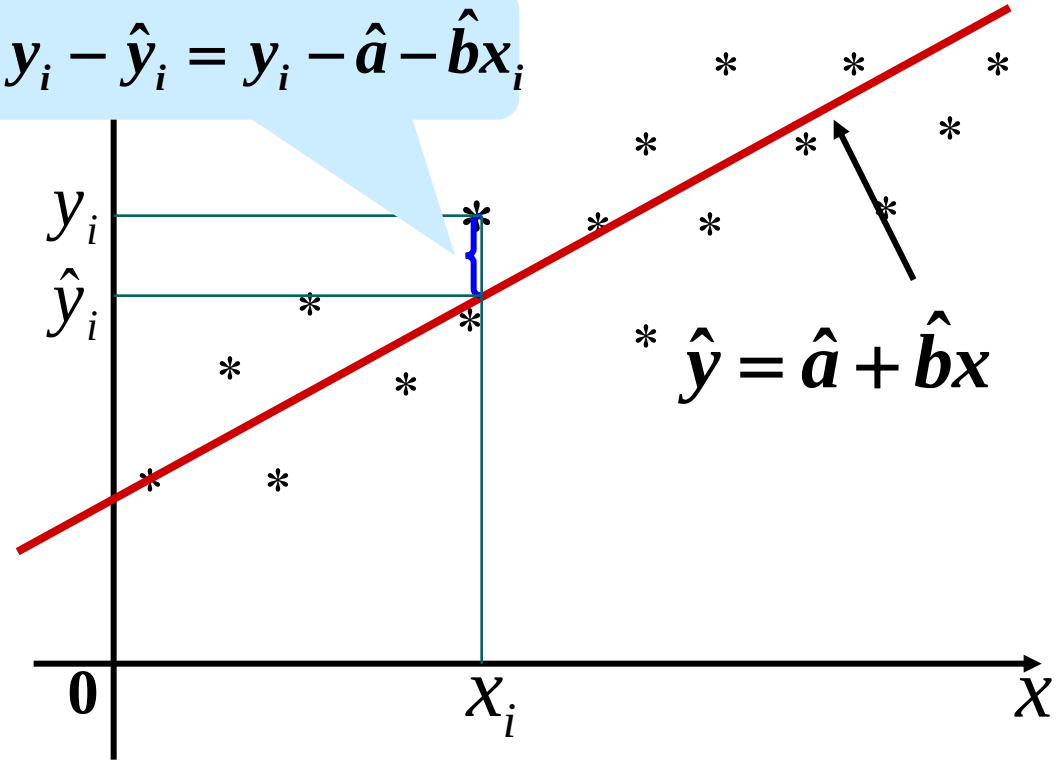
$$\begin{cases} \hat{a} = \bar{y} - \hat{b}\bar{x} \\ \hat{b} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \dfrac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \dfrac{l_{xy}}{l_{xx}} \end{cases}$$

回归方程： $\hat{y} = \hat{a} + \hat{b}x = \bar{y} + \hat{b}(x - \bar{x})$

记 $l_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, l_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2,$

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

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{a} - \hat{b}x_i$$



σ^2无偏估计

称$\hat{\varepsilon}_i = y_i - \hat{y}_i$为残差，$\hat{\varepsilon}_i$是$\varepsilon_i$的估计，

又 $\varepsilon_i \sim N(0,\sigma^2)$, 有 $D(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2$.

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

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

故，$s^2 = \dfrac{Q}{n-2} = \dfrac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$是$\sigma^2$的无偏估计.

定理1

(1) $\hat{b} \sim N\left(b, \dfrac{\sigma^2}{l_{xx}}\right)$;

(2) $\hat{a} \sim N\left[a, \sigma^2\left(\dfrac{1}{n}+\dfrac{\bar{x}^2}{l_{xx}}\right)\right]$;

(3) $\dfrac{(n-2)s^2}{\sigma^2} \sim \chi^2(n-2)$;

(4) $\bar{y}$, $\hat{b}$ 和 s^2 三者独立.

$$\hat{b}=\frac{l_{xy}}{l_{xx}},\quad \hat{a}=\bar{y}-\hat{b}\bar{x}$$

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

$$=\frac{1}{n-2}(l_{yy}-\hat{b}l_{xy})$$

一元线性回归模型

例1 根据《中国统计年鉴》2007年版，统计1978至2006年间用于科学研究的总费用与科研基建费数据如下：

8-7　国家财政用于科学研究的支出

单位：亿元

年 份	合 计	科研基建费
1978	52.89	6.66
1980	64.59	11.27
1985	102.59	18.83
1990	139.12	17.47
1991	160.69	18.40
1992	189.26	24.55
1993	225.61	33.95
1994	268.25	36.06
1995	302.36	38.00
1996	348.63	48.55
1997	408.86	42.74
1998	438.60	47.28
1999	543.85	52.89
2000	575.62	61.52
2001	703.26	63.37
2002	816.22	69.99
2003	975.54	111.06
2004	1095.34	95.90
2005	1334.91	112.50
2006	1688.50	134.40

注：表中科研基建费由国家统计局测算，2006年数据为初步数

例1(续) 回归方程：$\hat{y} = 18.157 + 0.0717x$

方差分析

源	自由度	平方和	均方	F 统计量	Pr > F
模型	1	9286.2449	9286.2449	118.70	<.0001
误差	9	704.1230	78.2359		
C 合计	10	9990.3680			

方差σ^2的估计s^2

参数估计值

变量	自由度	估计值	标准误差	T 统计量	Pr >|t|	容差	方差膨胀因子 (VIF)
Intercept	1	18.1570	5.9728	3.04	0.0140	.	0
all	1	0.0717	0.0066	10.89	<.0001	1.0000	1.0000

$\hat{a}$ $\hat{b}$

用SAS算的结果

回归方程的显著性检验

在实际中，事先我们并不能判定 X 与Y 确有线性关系，只是一种假设. 这种假设是根据专业知识和散点图作出的粗略判断. 在求出回归方程之后，还需对回归方程同实际观测数据拟合的效果进行检验.

回归方程的显著性检验

检验假设：$H_0: b=0,\quad H_1: b\neq 0$

若拒绝H_0,说明回归方程是显著线性的；

若接受H_0,说明Y与x间不是线性的，回归方程无意义，原因可能为：

(1) 影响Y的因素除x外，还有其他重要因素；

(2) Y对x不能用线性关系表示；

(3) Y与x无关系.

回归方程的显著性检验

检验假设：$H_0: b=0,\quad H_1: b\neq 0$ 有3种方法：

方法1 F检验法（也称方差分析法）

用方差分析法：

令 $S_T = l_{yy} = \sum_{i=1}^{n}(y_i-\bar{y})^2$ —— 总平方和

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

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

回归方程的显著性检验

总平方和分解：$S_T = Q + U$

$$S_T = l_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$ —总平方和

$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$ —残差平方和或剩余平方和

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

回归方程的显著性检验

定理2 当$b=0$时，

$$\frac{Q}{\sigma^2}=\frac{(n-2)s^2}{\sigma^2}\sim\chi^2(n-2)$$

(1) $\dfrac{S_T}{\sigma^2}=\dfrac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\sigma^2}\sim\chi^2(n-1)$;

(2) $\dfrac{U}{\sigma^2}=\left(\dfrac{\hat{b}\sqrt{l_{xx}}}{\sigma}\right)^2\sim\chi^2(1)$;

(3) $Q=(n-2)s^2$与U相互独立；

(4) $F=\dfrac{U}{Q/(n-2)}\sim F(1,n-2)$. ⇐检验统计量

回归方程的显著性检验

方差分析表

方差来源	平方和	自由度	均方	F值
回归	U	1	U	$\dfrac{U}{Q/(n-2)}$
剩余	Q	$n-2$	$Q/(n-2)$	
总和	S_T	$n-1$		

对给定显著性水平α，当$F \geq F_\alpha(1,n-2)$时，拒绝H_0，说明回归方程是显著线性的.

回归方程的显著性检验

方法2 t检验法

由定理1知 $\hat{b} \sim N\left(b, \dfrac{\sigma^2}{l_{xx}}\right) \Rightarrow \dfrac{\hat{b}-b}{\sigma / \sqrt{l_{xx}}} \sim N(0,1)$

$$\frac{(n-2)s^2}{\sigma^2} \sim \chi^2(n-2)$$

当$H_0: b=0$成立时，$t = \dfrac{\hat{b}}{s}\sqrt{l_{xx}} \sim t(n-2)$

对给定显著性水平α，当$|t| \geq t_{\alpha/2}(n-2)$时，

拒绝H_0，否则接受H_0.

回归方程的显著性检验

例1(续) 回归方程：$\hat{y} = 18.157 + 0.0717x$

方差分析

源	自由度	平方和	均方	F 统计量	Pr > F
模型	1	9286.2449	9286.2449	118.70	<.0001
误差	9	704.1230	78.2359		
C 合计	10	9990.3680			

显著

s^2

参数估计值

变量	自由度	估计值	标准误差	T 统计量	Pr >|t|	容差	方差膨胀因子 (VIF)
Intercept	1	18.1570	5.9728	3.04	0.0140	.	0
all	1	0.0717	0.0066	10.89	<.0001	1.0000	1.0000

$\hat{a}$　　$\hat{b}$

显著

回归方程的显著性检验

方法3 r检验法

$$\frac{U}{S_T}=\frac{\hat{b}^2 l_{xx}}{l_{yy}}=\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}\right]^2\overset{\Delta}{=}r^2$$

称r为样本相关系数.

回归方程的显著性检验

$$r^2=\frac{U}{S_T}=\frac{S_T-Q}{S_T}=1-\frac{Q}{S_T}\Rightarrow Q=S_T(1-r^2)$$

说明$|r|\le 1$,当S_T给定时，$|r|$越接近1，Q 越小，

回归方程线性关系越显著.

检验统计量 $$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

对给定α，查临界值$r_\alpha(n-2)$,当$|r|\ge r_\alpha(n-2)$时，

拒绝H_0,说明回归方程是显著线性的.

利用回归方程进行预测

预测：对给定的x值，预测Y的值.

对$Y=a+bx+\varepsilon$, 若回归方程$\hat{y}=\hat{a}+\hat{b}x$是显著的，可用于Y值的预测.

(1) 点估计值：对给定的x_0,可得y_0的预测值

$\hat{y}_0=\hat{a}+\hat{b}x_0$.

利用回归方程进行预测

(2) 区间估计：对给定的x_0，可以证明

$$t=\frac{y_0-\hat{y}_0}{s\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}}\sim t(n-2)$$

其中$s=\sqrt{Q/(n-2)}$称为剩余标准差.

给定$\alpha(0<\alpha<1)$, 有

$$P\{|t|<t_{\alpha/2}(n-2)\}=1-\alpha,$$

利用回归方程进行预测

对任意$x_0, y_0 = a + bx_0 + \varepsilon_0$的置信度为$1-\alpha$的

预测区间为 $(\hat{y}_0 - \delta(x_0), \hat{y}_0 + \delta(x_0))$,

其中$\hat{y}_0 = \hat{a} + \hat{b}x_0$,

$$\delta(x_0) = t_{\alpha/2}(n-2)s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}}.$$

利用回归方程进行控制

控制问题是使$Y=a+bx+\varepsilon$的值以$1-\alpha$的概率落在指定区间(y',y'')内时，求回归变量x应控制的范围问题.

给定置信度$1-\alpha$, 由预测区间

$$P\{\hat{y}-\delta(x)<y<\hat{y}+\delta(x)\}=1-\alpha$$

解不等式
$$\begin{cases}\hat{y}-\delta(x)\geq y',\\ \hat{y}+\delta(x)\leq y''.\end{cases}$$

如果不等式有解，即得回归变量x 的控制范围.

课程全部结束，衷心感谢大家的支持与帮助！愿大家有所收获.