§9.2 一元回归分析

本节讨论回归函数是一元线性函数和可线性化函数的情况.

一. 一元线性回归模型

下面我们对最简单的情形：一个自变量、回归函数是线性的一元线性回归方程进行讨论。

深度学习：
涉及线性回归模型，处理回归问题；
logistic模型处理分类问题。

若Y关于X 的回归函数为

$$\mu(x) = E(Y \mid X = x) = a + bx$$

有一元线性回归模型：

$Y=a+bx+\varepsilon$, **其中a、b为未知参数，且**

a — 回归常数(又称截距)

b — 回归系数(又称斜率)

ε —随机误差(随机扰动项)

若随机误差$\varepsilon \sim N(0,\sigma^2)$, 称为

一元线性正态回归模型, **其中**a,b,σ^2为未知参数.

$Y=a+bx+\varepsilon$ （一元线性正态回归模型）

取定自变量X 的一组值: $x_1, x_2, \ldots, x_n$,

对Y 做 n 次独立试验, 一组样本记为

$$Y_1, Y_2, \ldots, Y_n$$

则有 $Y_i = \underbrace{a + bx_i}_{\text{由自变量}X\text{确定的成分}} + \varepsilon_i$, $i= 1,\ldots,n$

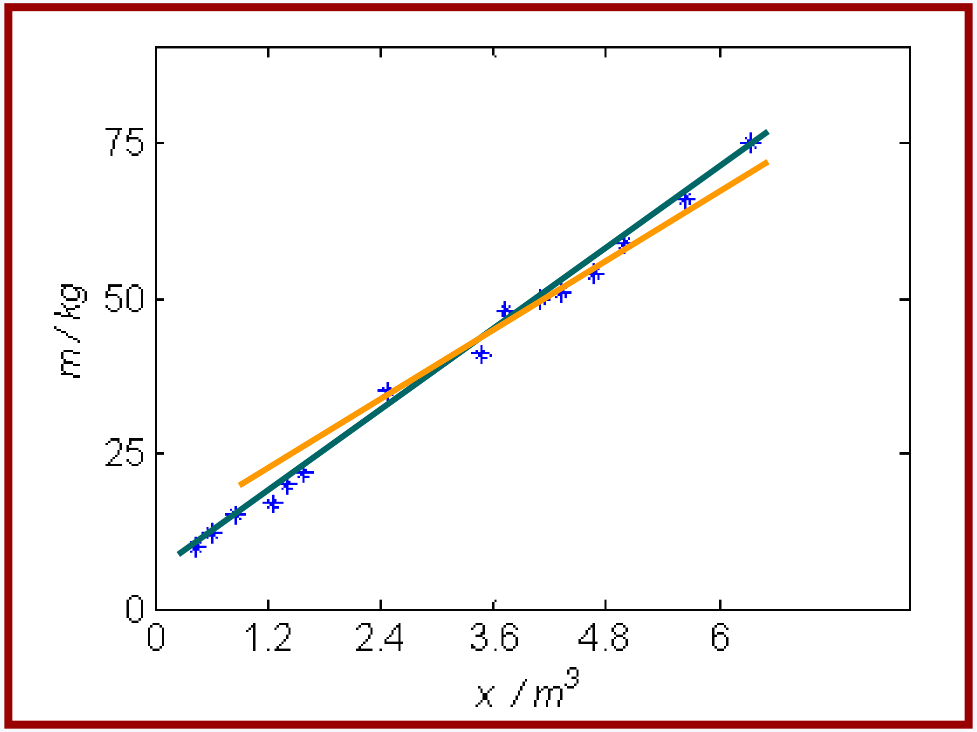
ε_i是第 i 次观察时的随机误差，有

回归假定

$$\begin{cases} 1)\ E(\varepsilon_i)=0,\ D(\varepsilon_i)=\sigma^2, i=1, 2, \ldots, n; \\ 2)\ \varepsilon_1, \ldots, \varepsilon_n \text{相互独立}. \end{cases}$$

二.一元线性回归模型的参数估计(a, b, σ^2)

对自变量X的一组值$x_1, x_2, \ldots, x_n$做n次独立试验，得独立观察值$y_1, y_2, \ldots, y_n$.



问题 如何依据观察值$(x_i, \ y_i), i=1,2,\ldots, n.$

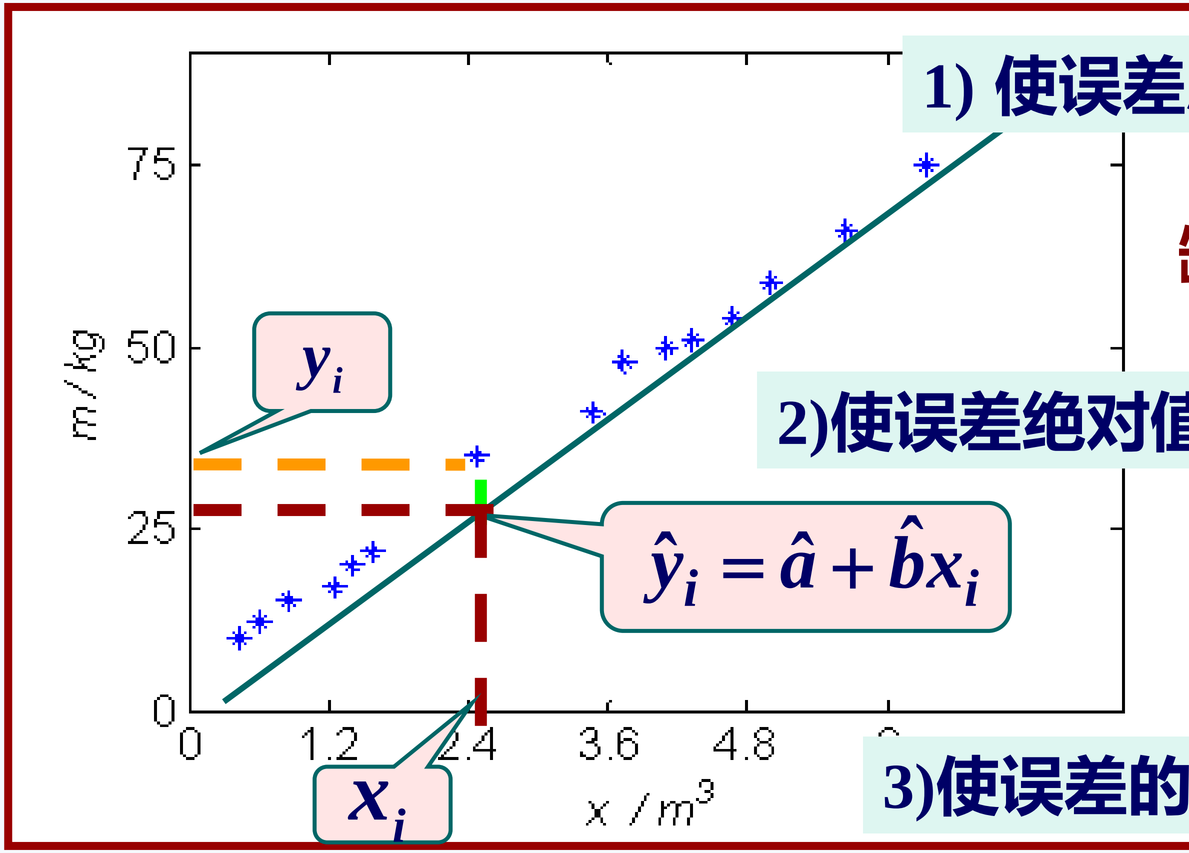
求a、b 的估计值 $\hat{a}, \hat{b}$？

记y_i 的估计值为

$$\hat{y}_i = \hat{a} + \hat{b}x_i$$

称为回归值.

对所有的i，应使偏差$y_i - \hat{y}_i$都尽可能小,有三种思路：



1) 使误差总和 $\sum(y_i - \hat{y}_i)$最小;

缺点:可能正负误差抵消

2)使误差绝对值之和 $\sum\left|y_i - \hat{y}_i\right|$最小;

缺点：数学处理困难

3)使误差的平方和$\sum(y_i - \hat{y}_i)^2$最小.

结论: 应选a、b的估计使残差(误差)平方和:

$$Q = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2$$ 达最小.

结论: 应选a、b的估计使 残差(误差)平方和:

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

应用最小二乘法: 使Q对应于a、b的一阶偏导为0,

$$\begin{cases}\dfrac{\partial Q}{\partial a}=-2\sum(y_i-a-bx_i)=0\\ \dfrac{\partial Q}{\partial b}=-2\sum(y_i-a-bx_i)x_i=0\end{cases}$$

化简为: $$\begin{cases}na+b\sum x_i=\sum y_i & (1)\\ a\sum x_i+b\sum x_i^2=\sum x_iy_i & (2)\end{cases}$$

$(2)\times n-(1)\times\sum x_i$得:

$$b[n\sum x_i^2-(\sum x_i)^2]=n\sum x_iy_i-\sum x_i\sum y_i$$

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

则：$\hat{b}=\dfrac{n\sum x_iy_i-\sum x_i\sum y_i}{n\sum x_i^2-(\sum x_i)^2}=\dfrac{\sum x_iy_i-n\bar{x}\bar{y}}{\sum x_i^{\ 2}-n\bar{x}^2}$

$$=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}=\frac{l_{xy}}{l_{xx}}$$

其中,$l_{xx}=\sum(x_i-\bar{x})^2$,

$l_{xy}=\sum(x_i-\bar{x})(y_i-\bar{y})$

由(1)：$\hat{a}=\dfrac{\sum y_i}{n}-\hat{b}\dfrac{\sum x_i}{n}=\bar{y}-\hat{b}\bar{x}$

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

$$\begin{cases}\hat{b}=\dfrac{l_{xy}}{l_{xx}}\\ \hat{a}=\bar{y}-\hat{b}\bar{x}\end{cases}$$

$\hat{a}$、$\hat{b}$分别是a、b的最小二乘估计值

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

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

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

由回归假定：$E(\varepsilon_i)=0$，$D(\varepsilon_i)=\sigma^2$，$i=1, 2, \ldots, n$;

$\sigma^2=D(\varepsilon)=E(\varepsilon^2)$，故 $\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2$ 是σ^2的矩估计量.

可证明σ^2的无偏估计量为

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

代入 $\hat{y}_i=\hat{a}+\hat{b}x_i$得：$\hat{\sigma}^2=\frac{1}{n-2}(l_{yy}-\hat{b}^2 l_{xx})$

其中：$l_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}y_i^2-n\bar{y}^2$ (教材: (9.2.6))

$$l_{xx}=\sum(x_i-\bar{x})^2=\sum x_i^2-n\bar{x}^2$$

$$l_{xy}=\sum(x_i-\bar{x})(y_i-\bar{y})=\sum x_i y_i-n\bar{x}\cdot\bar{y}$$

$$l_{yy}=\sum(y_i-\bar{y})^2=\sum y_i^2-n\bar{y}^2$$

最小二乘法：

$$\hat{b}=\frac{l_{xy}}{l_{xx}}$$

$$\hat{a}=\bar{y}-\hat{b}\bar{x}$$

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

Y对X的经验回归直线方程：

$$\hat{y}=\hat{a}+\hat{b}x$$

$$\neq y=a+bx$$

(理论)回归直线方程

可证：$E(\hat{a})=a,\ E(\hat{b})=b, E(\hat{\sigma}^2)=\sigma^2$(见教材)

例1第一导丝盘速度Y是合成纤维抽丝的重要因素，它与电流周波X(单位：HZ)有密切关系，由生产记录得10对数据：

$$\bar{x}=\frac{1}{10}\sum_{i=1}^{10}x_i=49.61 \qquad \sum_{i=1}^{10}x_i^2=24613.51$$

$$\bar{y}=\frac{1}{10}\sum_{i=1}^{10}y_i=16.86 \qquad \sum_{i=1}^{10}y_i^2=2842.84 \qquad \sum_{i=1}^{10}x_iy_i=8364.92$$

试求：Y对X的经验回归直线方程，并求误差方差σ^2的无偏估计值$\hat{\sigma}^2$，估计电流周波为50.5时丝盘速度为多少？

解：$l_{xx}=\sum_{i=1}^{10}x_i^2-n\bar{x}^2=1.989$ $\qquad l_{xy}=\sum_{i=1}^{10}x_iy_i-n\bar{x}\bar{y}=0.674$

$$l_{yy}=\sum_{i=1}^{10}y_i^2-n\bar{y}^2=0.244$$

例9.2.1(教材)

$$l_{xx}=\sum_{i=1}^{10}x_i^2-n\overline{x}^2=1.989$$

$$l_{xy}=\sum_{i=1}^{10}x_iy_i-n\overline{x}\,\overline{y}=0.674$$

$$\hat{b}=\frac{l_{xy}}{l_{xx}}\approx 0.3389$$

$$l_{yy}=\sum_{i=1}^{10}y_i^2-n\overline{y}^2=0.244$$

$$\hat{a}=\overline{y}-\hat{b}\overline{x}\approx 0.0472$$

$$\hat{\sigma}^2=\frac{1}{n-2}\left(l_{yy}-\hat{b}^2 l_{xx}\right)\approx 0.0019$$

经验回归直线方程为： $\hat{y}=\hat{a}+\hat{b}x=0.0472+0.3389x$

电流周波为50.5时丝盘速度值为：

$$\hat{y}(50.5)=0.0472+0.3389\cdot 50.5=17.16165$$

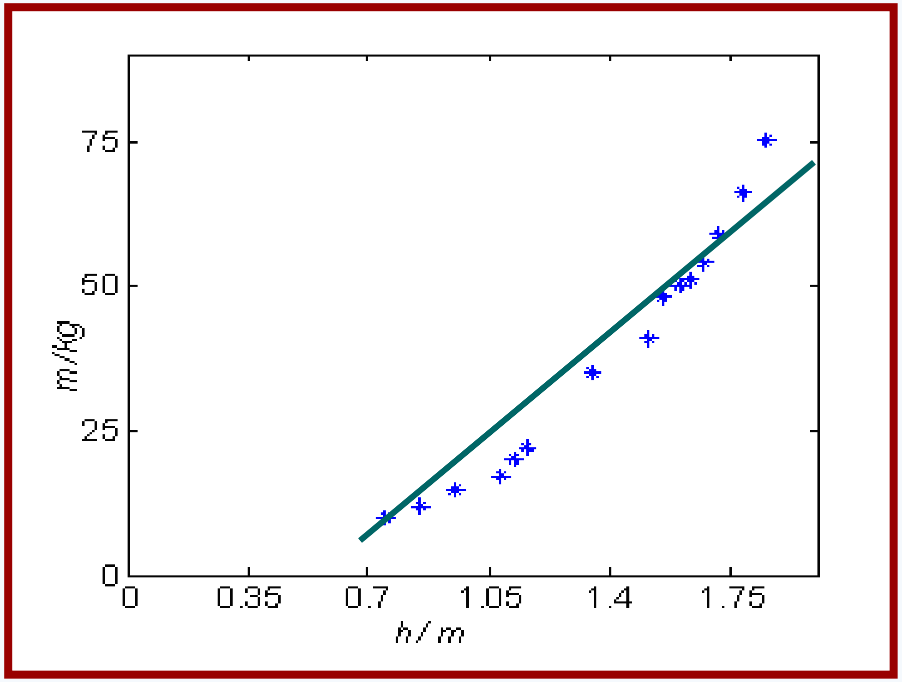
问题

随机变量Y与X间是否存在线性相关关系?

是否能由数据散布图完全确定回归函数?

续 例9.1.1

身高体重关系



身高h 和体重m无明显的线性相关关系.

形式地建立经验线性回归方程无实际意义.

三、一元线性回归的假设检验

1.相关系数法

基于试验数据检验随机变量间线性相关关系是否显著的一种方法.

相关系数

$$\rho_{XY}=\frac{E\{[X-E(Y)][Y-E(Y)]\}}{\sqrt{D(X)}\sqrt{D(Y)}}$$

表征随机变量Y与X的线性相关程度的数字特征.

$$\rho_{XY}=\frac{E\{[X-E(X)][Y-E(Y)]\}}{\sqrt{D(X)}\sqrt{D(Y)}}\xrightarrow{\text{实际中未知}}$$

用矩法估计值来替换

样本相关系数:

定义:

$$\hat{\rho}_{XY}=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$=\frac{l_{xy}}{\sqrt{l_{xx}}\sqrt{l_{yy}}}=R$$

作为ρ_{XY}的估计值.

补充: 两个向量的夹角的余弦,叫做“相关系数”

$\cos<a,b>=(a\cdot b)/|a|\cdot|b|$ **写开了就是上式.**

$$|R| = \left| \frac{l_{xy}}{\sqrt{l_{xx}}\sqrt{l_{yy}}} \right| \le 1 \text{(可证明)}$$

统计值R描述了X与Y间的线性相关关系的密切程度.

1) 当$R = 0$ $\longrightarrow$ $l_{xy} = 0$

$$\xrightarrow{\hat{y} = \hat{a} + \hat{b}x} \hat{b} = \frac{l_{xy}}{l_{xx}} = 0$$

$\longrightarrow$ 经验回归方程形为 $\hat{y} = \hat{a}$

说明自变量X的变化不会引起因变量Y的变化(不相关).

Y对X的经验回归直线方程：$\hat{y}=\hat{a}+\hat{b}x$

2）$|R|=1$ $\Rightarrow$ $\dfrac{l_{xy}}{\sqrt{l_{xx}}\sqrt{l_{yy}}}=\pm 1$

$\Rightarrow$ $l_{xy}=\pm\sqrt{l_{xx}}\sqrt{l_{yy}}$

$\Rightarrow$ $\hat{b}=\dfrac{l_{xy}}{l_{xx}}=\dfrac{\pm\sqrt{l_{xx}}\sqrt{l_{yy}}}{l_{xx}}=\pm\sqrt{\dfrac{l_{yy}}{l_{xx}}}\neq 0$

可认为X与Y间存在线性相关关系.

结论: 1) $|R|$ 越接近于1，X与Y间的线性相关关系越显著;

2) $|R|$ 越靠近于0，X与Y间的线性相关关系越不显著.

根据P306附表6 相关系数临界值表，有

判别准则 给定显著性水平α（0.05，0.01）

当 $|R| > R_\alpha(n-2)$,
认为X与Y之间的线性相关关系显著

当 $|R| \le R_\alpha(n-2)$,
认为X与Y之间的线性相关关系不显著

例2： 试分别用F检验法和相关系数检验法，检验例1中的电流周波X与第一导丝盘速度Y之间是否存在显著的线性相关关系？ (取α=0.01)

教材例9.2.2 (212页)

解: (1)相关系数检验法

$$H_0:\ \rho_{XY}=0$$

$$|R_{XY}|=\left|\frac{l_{xy}}{\sqrt{l_{xx}l_{yy}}}\right|=\frac{0.674}{\sqrt{1.989\cdot 0.244}}\approx 0.9675$$

$n=10$, 查表得$R_{0.01}(8)=0.765$

因$|R|>R_{0.01}(8)$ 拒绝接受$H_0:\ \rho_{XY}=0$

即这说明电流周波X与第一导丝盘速度Y之间线性相关关系显著.

第九章 回归分析(包含丢分或掉坑的地方)

基本思想: 根据自变量X_1, X_2, $\cdots$ X_k与因变量Y的观察值去估计回归函数.

1. 一元线性回归模型(标准形式): $Y=a+bx+\varepsilon$, $\varepsilon\sim N(0,\sigma^2)$

(最小二乘法思想: 误差或残差平方和的极小值点)

$$Q=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-\hat{a}-\hat{b}x_i)^2=\sum_{i=1}^{n}\varepsilon_i^2$$ (注意:书后的第1题)

(回归参数b,回归常数a的估计:重点看教材例9.2.1和例9.2.2)

一元经验线性回归方程: $\hat{y}=\hat{a}+\hat{b}x$ $\qquad \hat{b}=\dfrac{l_{xy}}{l_{xx}}$ $\quad \hat{a}=\bar{y}-\hat{b}\bar{x}$

不可少

2. (重点)相关系数的显著性检验法. $\hat{\sigma}^2=\dfrac{1}{n-2}(l_{yy}-\hat{b}^2 l_{xx})$

拒绝域: $|R|=|\hat{\rho}_{XY}|=\left|\dfrac{l_{XY}}{\sqrt{l_{XX}}\sqrt{l_{YY}}}\right|>R_\alpha(n-2)$ 例6

3.非线性回归问题的线性化处理.例9.4.3(教材)

回归分析的作用: (在生产实践上)预测和控制.

2. *F*检验法(自学)

四、非线性回归问题的线性化处理

1.用变量替换将非线性回归问题线性化

例9.4.3 对自变量实施变量替换：

对于对数回归方程如：$y=a+b\ln x$

变量代换：$x^*=\ln x$，将对应数据转换

线性回归方程为：$y=a+bx^*$

检验自变量X与因变量Y是否存在显著对数相关关系 ⇔

检验自变量$\ln X$与因变量Y是否存在显著线性相关关系

四. 非线性回归问题的线性化处理

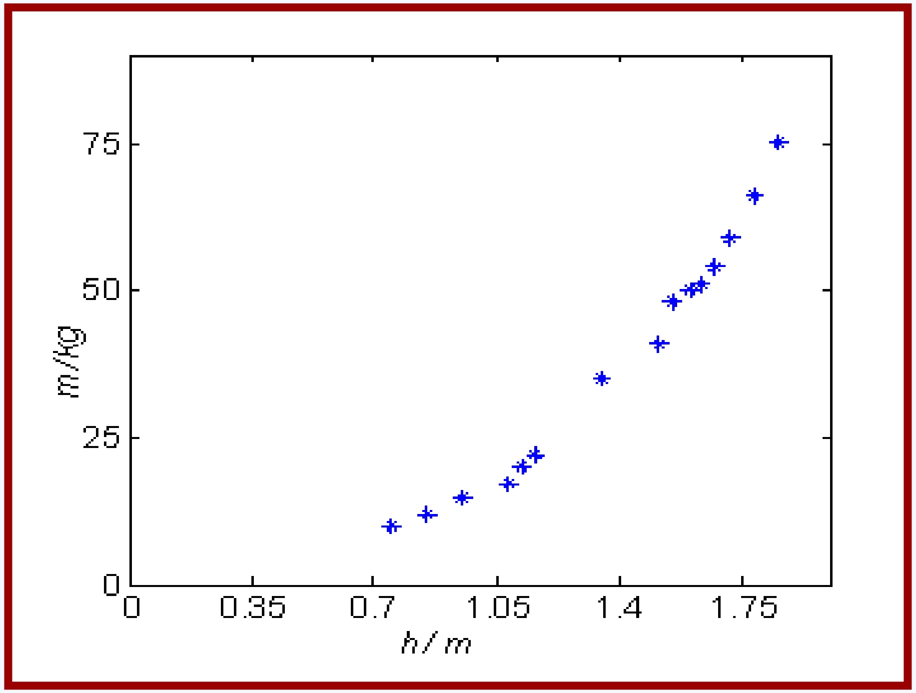
在实际问题中,变量间的相关关系未必是线性关系，即其回归函数

$y = \mu(x_1, x_2, \cdots, x_k)$ 往往是非线性函数.

可通过适当的变换，将其转化为线性回归问题.

例9.2.4建模范例（身高与体重关系）

现有15对某地区人的身高h 和体重数据m，希望用简洁的函数关系式描述该地区人的身高体重的对应关系.



对选定的回归函数的估计函数

$$m = dh^a$$

两边取对数得

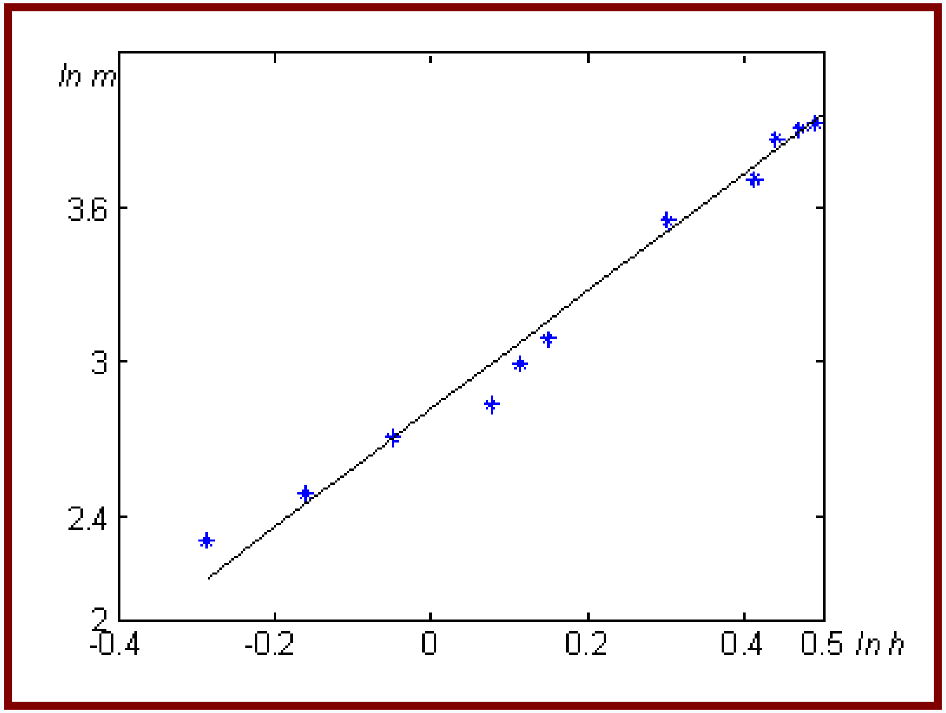
$$\ln m = a\ln h + \ln d$$

记 $y = \ln m, x = \ln h, b = \ln d$ 得线性方程

$$y = ax + b, b \neq 0$$

去掉m=0, h=0, 对原数据做相应变换.

变换数据的散布图



用*matlab*软件编程可得经验线性回归方程

即 $$\ln \hat{m} = 2.30 \ln h + 2.28$$

从而 $$\hat{m} = 16.78 h^{2.3}$$

即 $$\ln \hat{m} = 2.30 \ln h + 2.28$$

从而 $$\hat{m} = 16.78 h^{2.3}$$

可用来描述该地区人的身高和体重的关系.

问题 请考虑还需做什么工作?

检验h与m是否存在显著的幂函数相关关系

检验ln$h$与ln$m$的线性相关关系是否显著

相关系数法基于试验数据检验变量间线性相关系数是否显著的一种方法。补充例

例题9.2.1：炼钢是气化脱碳的过程．精炼时间是指炉料 熔毕至出钢所需要的时间，含碳量是指全部炉料熔化完毕时钢液的合碳量．钢液含碳量X与精炼时间Y有一定关联的关系，现将某平炉共炼$n＝34$炉钢记录下来的X与Y的数据列如下表．

试写出Y对X的经验回归直线方程，并求出误差方差的无偏估计值。并检验含碳量X与精炼时间Y之间是否存在显著的线性相关关系？$(\alpha=0.01)$

编号	含碳量 x(%)	精炼时间 y(分)	编号	含碳量 x(%)	精炼时间 y(分)
1	1.80	200	18	1.16	100
2	1.04	100	19	1.23	110
3	1.34	135	20	1.51	180
4	1.41	125	21	1.10	130
5	2.04	235	22	1.08	110
6	1.50	170	23	1.58	130
7	1.20	125	24	1.07	115
8	1.51	135	25	1.80	240
9	1.47	155	26	1.27	135
10	1.45	165	27	1.15	120
11	1.41	135	28	1.91	205
12	1.44	160	29	1.90	220
13	1.90	190	30	1.63	145
14	1.90	210	31	1.55	160
15	1.61	145	32	1.77	185
16	1.65	195	33	1.77	205
17	1.54	150	34	1.43	160

解： $\bar{x}=1.5009 \quad \sum_{i=1}^{34} x_i^2=79.1384$

$$\bar{y}=158.235 \quad \sum_{i=1}^{34} y_i^2=801201 \quad \sum_{i=1}^{34} x_i y_i=8389.81$$

$$l_{xx}=\sum_{i=1}^{34} x_i^2-34\bar{x}^2=2.5466$$

$$l_{xy}=\sum_{i=1}^{34} x_i y_i-34\bar{x}\,\bar{y}=323.2540$$

$$l_{yy}=\sum_{i=1}^{34} y_i^2-34\bar{y}^2=50094.002$$

$$\hat{b}=\frac{l_{xy}}{l_{xx}}=126.936 \qquad \hat{a}=\bar{y}-\hat{b}\,\bar{x}=-32.283$$

经验回归直线方程为：

$$\hat{y} = \hat{a} + \hat{b}\,x = -32.283 + 126.936x$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\left(l_{yy} - \hat{b}^2 l_{xx}\right)$$

$$= \frac{1}{34-2}\left(50094.002 - 126.936^2 * 2.5466\right) = 283.165$$

X, Y 的线性相关程度如何呢？

$$|R_{XY}| = \left|\frac{l_{xy}}{\sqrt{l_{xx} l_{yy}}}\right| \approx 0.9054 > R_{0.01}(30) = 0.449$$

$$> R_{0.01}(32)$$

拒绝接受H_0：$\rho_{XY} = 0$

这说明X与Y的线性相关关系显著.

(2) F检验法

$$Q_R = \hat{b}^2 l_{xx} = 41032.7243$$

$$Q_E = l_{yy} - \hat{b}^2 l_{xx} = 50094.002 - 41032.7243 = 9061.2777$$

$$F = \frac{Q_R}{Q_E/(n-2)} = \frac{41032.7243}{9061.2777/32} \approx 144.90751$$

$$F > F_{0.01}(1,30) = 7.56 > F_{0.01}(1,32)$$

拒绝接受H_0：$b=0$

这说明X与Y的线性相关关系显著.

例2及9.2.2(209页）

补充：

$$R^2 = \frac{l_{xy}^2}{l_{xx} l_{yy}} = \frac{\hat{b}^2 l_{xx}^2}{l_{xx} l_{yy}} = \frac{\hat{b}^2 l_{xx}}{l_{yy}} = \frac{Q_R}{Q_T}$$

$$\because Q_E = Q_T - Q_R = \frac{1}{R^2} Q_R - Q_R$$

$$F = \frac{Q_R / 1}{Q_E / (n-2)} = \frac{Q_R}{\frac{1-R^2}{R^2} Q_R}(n-2)$$

$$= \frac{(n-2)R^2}{1-R^2} \sim F(1, n-2)$$

$$\Rightarrow R_\alpha(n-2) = \sqrt{\frac{F_\alpha(1, n-2)}{(n-2) + F_\alpha(1, n-2)}}$$

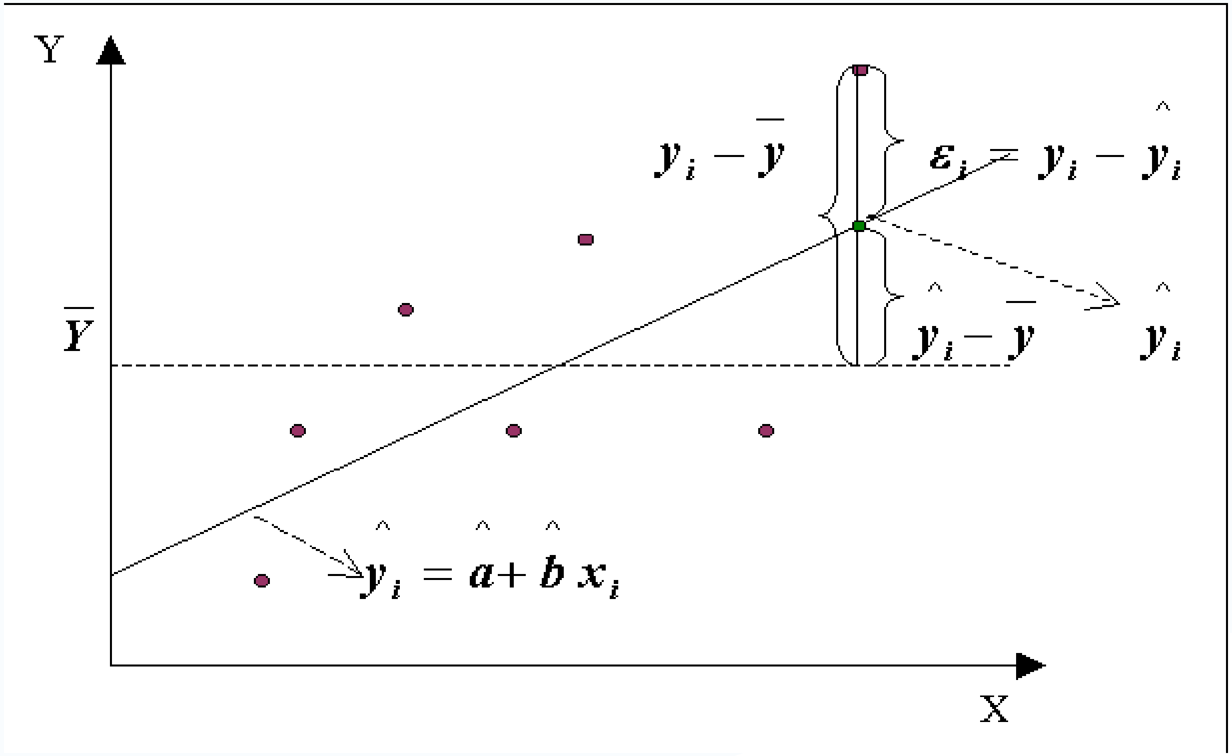
二、一元线性回归中的假设检验：

1. 基于总离差平方和分解的检验

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

书206

证明



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

$$Q_T = Q_E + Q_R$$

总离差平方和 ＝ 残差平方和 ＋ 回归离差平方和

不能由回归直线作出解释的部分，由随机误差影响

由回归直线作出解释的部分，自变量影响

直观意义：若$Q_R>Q_E$则说明拟合好，X、Y有显著相关关系； 否则说明X、Y无显著线性相关关系或不相关。

1、F检验法

依据：X、Y线性相关关系是否显著～b是否等于0

原假设： H_0：$b=0$

变差来源	总平方和 Q_T	回归平方和 Q_R	剩余平方和 Q_E
平方和	$\sum_{i=1}^{n}(y_i-\bar{y})^2$	$\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2$	$\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2$
自由度	n-1	1	n-2

自由度的说明详见“方差分析”

检验统计量：$$F=\frac{Q_R}{Q_E/(n-2)}\sim F(1,n-2)$$

原假设：H_0： $b=0$

检验统计量：$F=\dfrac{Q_R}{Q_E/(n-2)}\sim F(1,n-2)$

其中：$Q_T=l_{yy}$

$$Q_R=\sum\left(\hat{y}_i-\bar{y}\right)^2=\hat{b}^2 l_{xx},\quad Q_E=\sum\left(y_i-\hat{y}_i\right)^2=l_{yy}-\hat{b}^2 l_{xx}$$

拒绝域：$F=\dfrac{Q_R}{Q_E/(n-2)}>F_\alpha(1,n-2)$

如果拒绝H_0，则自变量X与因变量Y间线性相关关系显著；

如果接受H_0，则意味着回归函数为一常数，X与Y没有明显的线性相关关系。