$\mu(x)=a+bx$ 一元线性回归问题

假设对于x的每一个值有$Y\sim N(a+bx,\sigma^2)$,a,b,σ^2都是不依赖于x的未知参数.

记$\varepsilon=Y-(a+bx)$,那么

$Y=a+bx+\varepsilon,\quad \varepsilon\sim N(0,\sigma^2).$

a,b,σ^2是不依赖于x的未知参数.

一元线性回归模型

x的线性函数　　随机误差

未知参数a,b的估计

$Y=a+bx+\varepsilon,\quad \varepsilon\sim N(0,\sigma^2).$

对于样本 $(x_1,Y_1),(x_2,Y_2),\cdots,(x_n,Y_n)$

$Y_i=a+bx_i+\varepsilon_i,\ \varepsilon_i\sim N(0,\sigma^2)$, 各 ε_i 相互独立.

于是 $Y_i\sim N(a+bx_i,\sigma^2)$, $i=1,2,\cdots,n$.

根据$Y_1,Y_2,\cdots,Y_n$的独立性可得到联合密度函数为

$$L=\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}(y_i-a-bx_i)^2\right]$$

$$=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-a-bx_i)^2\right].$$

用最大似然估计估计未知参数 a,b.

对于任意一组观察值 $y_1,y_2,\cdots,y_n$, 样本的似然函数为 $L=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-a-bx_i)^2\right]$

L取最大值等价于

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

取最小值.

$$\left.\begin{aligned}\frac{\partial Q}{\partial a}&=-2\sum_{i=1}^{n}(y_i-a-bx_i)=0\\ \frac{\partial Q}{\partial b}&=-2\sum_{i=1}^{n}(y_i-a-bx_i)x_i=0\end{aligned}\right\}$$

$$\left.\begin{aligned}na+\left(\sum_{i=1}^{n}x_i\right)b&=\sum_{i=1}^{n}y_i\\ \left(\sum_{i=1}^{n}x_i\right)a+\left(\sum_{i=1}^{n}x_i^2\right)b&=\sum_{i=1}^{n}x_iy_i\end{aligned}\right\}$$ 正规方程组

$$\begin{vmatrix}n & \sum_{i=1}^{n}x_i\\ \sum_{i=1}^{n}x_i & \sum_{i=1}^{n}x_i^2\end{vmatrix}\neq 0,$$

$$\hat{b}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2},$$

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

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

$\mu(x)=a+bx$

$\hat{\mu}(x)=\hat{a}+\hat{b}x$　Y 关于 x 的经验回归函数

$\hat{y}=\hat{a}+\hat{b}x$　　Y 关于 x 的经验回归方程

回归方程　回归直线

由于$\hat{a}=\bar{y}-\hat{b}\bar{x}$,

$\hat{y}=\bar{y}+\hat{b}(x-\bar{x})$,

回归直线通过散点图的几何中心$(\bar{x},\bar{y})$.

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

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

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

$\hat{b}=\frac{S_{xy}}{S_{xx}},\qquad \hat{a}=\frac{1}{n}\sum_{i=1}^{n}y_i-\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)\hat{b}.$

非线性回归与曲线回归

一、非线性回归模型的类型

（一）抛物线模型(二次曲线模型)

$$Y=\beta_0+\beta_1X+\beta_2X^2+\varepsilon$$

式中β_0、β_1和β_2为待估计参数。

（二）双曲线模型 $Y=\beta_0+\beta_1\ (1/X)\ +\varepsilon$

（三）幂函数模型

$$Y=\beta_0X_1^{\beta_1}X_2^{\beta_2}\cdots\cdots X_p^{\beta_p}e^{\varepsilon}$$

（四）指数函数模型 $Y=\beta_0e^{\beta_1X+\varepsilon}$

（五）对数函数模型 $Y=\beta_0+\beta_1\ln X+\varepsilon$

(六)逻辑曲线模型

$$Y=\frac{L}{1+\beta_0e^{-\beta_1X+\varepsilon}}\quad (L>0)$$

（七）多项式模型

$$Y=\beta_0+\beta_1X+\beta_2X^2+\cdots\beta_pX^p+\varepsilon$$

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_1X_2+\beta_4X_1^2+\beta_5X_2^2+\varepsilon$$

二，可化为一元线性回归的问题

方法——通过适当的变量变换,化成一元线性回归问题进行分析处理.

（一）倒数变换

$$Y=\beta_0+\beta_1\ (1/X)\ +\varepsilon$$

令 $X^*=1/X$

得 $Y=\beta_0+\beta_1X^*+\varepsilon$

（二）半对数变换

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

令 $X^*=\ln X$，可得：$Y=\beta_0+\beta_1X^*+\varepsilon$

（三）双对数变换

$$Y=\beta_0X_1^{\beta_1}X_2^{\beta_2}\cdots\cdots X_p^{\beta_p}e^{\varepsilon}$$

两边求对数，可得：

$$\ln Y=\ln\beta_0+\beta_1\ln X_1+\beta_2\ln X_2+\cdots+\beta_p\ln X_p+\varepsilon$$

令 $Y^*=\ln Y$；$\beta_0^*=\ln\beta_0$;$X_1^*=\ln X_1,\cdots,\ X_k^*=\ln X_k$,

可得：$Y^*=\beta_0^*+\beta_0X_1^*+\beta_1X_2^*+\cdots+\beta_pX_p^*$

（四）多项式变换

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_1X_2+\beta_4X_1^2+\beta_5X_2^2+\varepsilon$$

令 $X_2^*=X_1,X_3^*=X_2,X_4^*=X_1X_2,X_5^*=X_1^2,X_6^*=X_2^2$

可得：

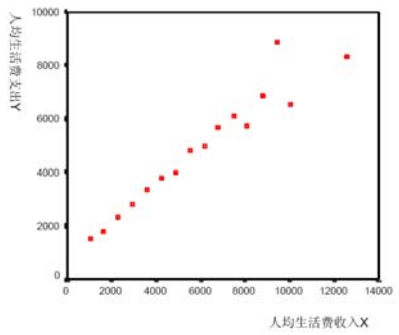
$$Y=\beta_0+\beta_1X_2^*+\beta_2X_3^*+\beta_3X_4^*+\beta_4X_5^*+\beta_5X_6^*$$

例：1996 年我国城镇居民收入情况如所表所示：表中资料共有16 组，X 是各组的人均生活费收入，Y 是各组的人均生活费支出。试建立Y 对X 的回归模型。

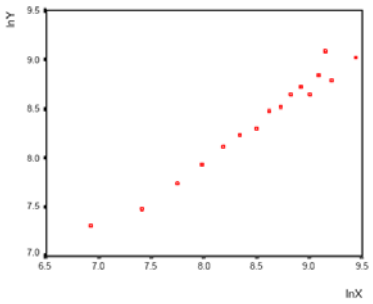
（单位：元）

Y	X	Y	X
1493.47	1017.52	4996.12	6160.77
1762.82	1643.86	5692.75	6785.27
2298.40	2300.04	6102.06	7503.47
2784.98	2917.52	5712.40	8106.87
3345.43	3567.42	6886.65	8814.28
3769.01	4205.01	8877.78	9427.21
3981.00	4881.93	6561.70	10001.90
4805.03	5521.33	8311.68	12582.52

从如下的散点图可以看出，人均生活费支出先是随着人均生活费收入的提高而快速提高，但当收入达到一定水平后，生活费支出的增幅明显趋缓。因此，用线性回归模型表示Y 和X 的关系是不恰当的。



对Y 和X 分别取自然对数lnY 和lnX，画出lnY 和lnX 的散点图。



可以看出，lnY 和lnX 在散点图上近似为线性关系。
于是把回归模型高定为幂函数模型：

$$Y = \beta_0 X^{\beta_1} e^{\varepsilon}$$

并进行双对数变换，得

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

分别令 $Y^* = \ln Y, \beta_0^* = \ln \beta_0, X^* = \ln X$，得到线性回归模型：

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

回归结果如下：

$$\hat{Y}^* = 2.006 + 0.748 X^*$$

t值: 7.098　22.580

p_t值: 0.000　0.000

$R^2 = 0.973$　$F = 509.847$　$p_F = 0.000$

模型通过参数显著性检验。注意到这里的 $\hat{\beta}_1$ 是弹性值，即人均生活费收入每提高1%，人均生活费支出平均增加0.748%。

三、曲线估计

非线性模型方程式

名　称		方程式
Linear（一元线性）	LIN	$Y=\beta_0+\beta_1 t$
Quadratic(二次函数)	QUA	$Y=\beta_0+\beta_1 t+\beta_2 t^2$
Compoound(复合函数)	COM	$Y=\beta_0(\beta_1)^t$
Growth(生长函数)	GRO	$Y=e^{(\beta_0+\beta_1 t)}$
Logarithmic(对数函数)	LOG	$Y=\beta_0+\beta_1 \ln t$
Cubic(三次函数)	CUB	$Y=\beta_0+\beta_1 t+\beta_2 t^2+\beta_3 t^3$
S(S 形曲线)	S	$Y=e^{(\beta_0+\beta_1/t)}$

Exponential(指数函数)	EXP	$Y=\beta_0 e^{\beta_1 t}$
Inverse(逆函数)	INV	$Y=\beta_0+\beta_1/t$
Power(幂函数)	POW	$Y=\beta_0 t^{\beta_1}$
Logistic(逻辑函数)	LGS	$Y=\frac{1}{\frac{1}{u}+\beta_0(\beta_1 t)}$

表中，t 为时间或自变量，β_0为常数项，β_1为回归参数，e 表示自然对数的底，ln 表示以 e为底的自然对数。

小结

1.回归分析的任务

研究变量之间的相关关系

2.一元线性回归的步骤

(1) 推测回归函数;　　(2) 建立回归模型;
(3) 估计未知参数;　　(4) 进行假设检验;
(5) 预测与控制.

3. 非线性回归，可化为一元线性回归的问题

4. 曲线回归