第一章　回归分析简介

1. 1 曲线拟合

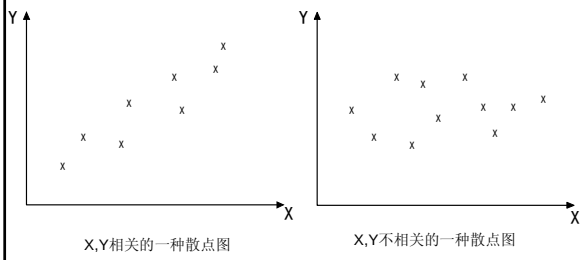
1.1.1 散点图

设有 n对观测值$(X_i,Y_i)\ i=1,2,\cdots n.$“确定”两个变量 X,Y 之间所存在的关系。

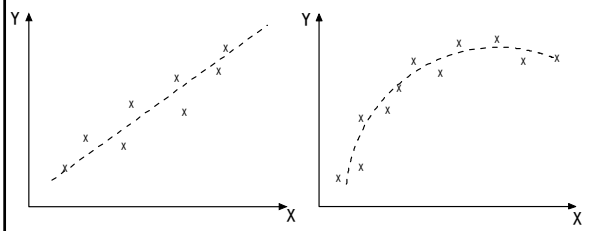
通常作法：

1.作散点图；

2.根据散点图尽量拟合一条“优美”曲线，使这些点尽可能“趋近”这条曲线。



X,Y相关的一种散点图　　X,Y不相关的一种散点图



X,Y呈线性关系　　X,Y呈非线性关系

1.1.2 拟合曲线的选择

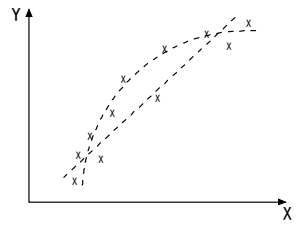
- 拟合一条“优美”的曲线通常指拟合一条光滑的曲线；
- 由于许多不可控因素使变量发生随机波动，不要期望点与曲线完全拟合；
- 即使两变量之间存在确定关系，由于测量误差，这种偶然性的波动仍然会在散点图上表现出来。

- 按经验规律或有关理论等选择拟合曲线的类型

例 1：设一组试验数据(X,Y)，其中X是通过某电阻的电流安培数，Y是加在该电阻的电压伏特数。由欧姆定律X,Y有关系$Y=r\cdot X$，过原点呈直线型的散点图可印证这个定律。用过原点的直线去拟合这组数据，还可以用直线的斜率去估计电阻 r的值。

例 2：对保持恒温的一定质量的气体得到一组试验数据(V,P)，根据物理学原理V,P有关系$PV^r=c$，取对数得到线性关系$\log P=\log c-r\log V$，因此$\log c$和 r可以由拟合试验数据$(\log V,\log P)$的直线估计出来。

- 当没有经验规律和有关理论帮助时，确定应该拟合曲线的类型有时是困难的。



拟合同一组数据的两条不同曲线

一般说来若假设拟合的曲线形状已知，只是包含若干未知参数，由已知数据来估计这些未知参数，此方法称为参数的方法；若对拟合的曲线形状没有假定，一般用非参数的方法来拟合曲线。

1.2 回归分析

1.2.1 协变量与响应变量

在现实世界中存在大量这样的情况：两个或多个变量之间有一些联系，但可能没有确切到可以严格决定的程度。

例 1：人的身高X与体重Y有联系，一般表现为X大时Y也倾向于大，但由X不能严格决定Y。

例 2：一种农作物的亩产量Y与其播种量X_1施肥量X_2有联系，但X_1, X_2不能严格决定Y。

以上例子以及类似的例子中，Y通常称为响应变量(response)(或因变量、预报变量)，X, X_1, X_2等则称为协变量(covariates)(或自变量、预报因子)。协变量可以是随机的，也可以是非随机的，通常是可以观测的。

1.2.2 回归函数

现设在一个问题中有响应变量Y，协变量$X_1, \cdots X_p$。可以设想响应Y由两部分构成：一部分由$X_1, \cdots X_p$的影响所致，这一部分可以表为$X_1, \cdots X_p$的函数形式$f(X_1, \cdots X_p)$；另一部分则由其他众多未加考虑的因素，包括随机因素的影响所致，这部分视为一种随机误差，记为e。

于是得到模型：

$$Y = f(X_1, \cdots X_p) + e$$

这里e作为随机误差要求$Ee = 0$。

上式也可写为

$$f(X_1, \cdots X_p) = E\left(Y \middle| X_1, \cdots X_p\right)$$

称为Y对$X_1, \cdots X_p$的理论回归函数(regression function)。回归函数前面所加“理论”两字是为区分由数据估计所得的回归函数(称为经验回归函数)。

在实际问题中，理论回归函数一般总是未知的，统计回归分析的任务在于根据 $X_1,\cdots X_p$ 和 Y 的观测值去估计回归函数及讨论与此有关的一些统计推断问题。所用的方法在很大程度上取决于对模型中回归函数 f 及随机误差 e 所作的假定。若对回归函数 f 的数学形式并无特殊假定，称为非参数回归(non-parametric regression)；若假定 f 的形式已知，只是其中若干参数未知，这种情况称为参数回归(parametric regression)。

一般说来 在参数回归中，若 f 关于未知参数是线性的，称为线性回归(linear regression)；若关于参数是非线性的，称为非线性回归(nonlinear regression)。

对于随机误差 e，已经假定其均值 $Ee=0$，其方差 $Var(e)=\sigma^2$ 是模型的一重要参数。由于

$$E\left(Y-f(X_1,\cdots X_p)\right)^2=Ee^2=\sigma^2,$$

因此 σ^2 越小，用回归函数 $f(X_1,\cdots X_p)$ 逼近 Y 的均方误差(mean square error)就越小。

误差方差 σ^2 的大小主要由以下两点决定：

1).选择协变量时，是否把对响应变量有重要影响的那些因素都包括了；

2).回归函数的形状是否选择准确。

1.3 线性回归模型与最小二乘法

1.3.1 线性回归模型(linear regression models)

设响应变量 y 与协变量 $x_1,\cdots x_p$ 有关系

$$y=x_1\beta_1+\cdots x_p\beta_p+e, Ee=0,$$

现有 n 次观测，即

$$y_i=x_{i1}\beta_1+\cdots x_{ip}\beta_p+e_i, i=1,\cdots,n.$$

这里 $(y_i,x_{i1},\cdots x_{ip})$ 为已知，$(\beta_1,\cdots,\beta_p)$ 为未知非随机的参数，e_i 是第 i 次观测的随机误差，且假定 $Ee_i=0$，$Cov(e_i,e_j)=\sigma_{ij}, 1\le i,j\le n.$

通常写成简洁的矩阵形式。令

$$X_{n\times p}=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

$$Y=(y_1,\cdots,y_n)', \beta=(\beta_1,\cdots,\beta_p)'$$

$$e=(e_1,\cdots,e_n)', \Sigma=Cov(e)=(\sigma_{ij})_{n\times n}.$$

这样得到矩阵表示：

$$\begin{cases} Y=X\beta+e, \\ Ee=0, Cov(e)=\Sigma. \end{cases} \quad (1)$$

模型(1)描述变量之间线性相关关系的一类统计模型，称为线性回归模型(注意关于未知参数 β 是线性的)。向量 Y 通常称为观测向量，矩阵 X 通常称为设计矩阵(design matrix)。

设计矩阵 X 的秩 $rank(X)$ 可以等于 p(此时 X 称为满秩的)，也可以小于 p。一般线性回归模型设计矩阵 X 是满秩的，方差分析模型和协方差分析模型的设计矩阵 X 不是满秩的。设计矩阵 X 是否为满秩，对模型(1)的参数估计问题会产生很大的影响。

1.3.2 GM 条件

对于误差协方差阵 $\Sigma = Cov(e)$，若为未知，则难以由 n 次观测数据去估计模型(1)的 p 个未知系数 $\beta_1, \cdots, \beta_p$ 以及 $n(n+1)/2$ 个未知量 $\sigma_{ij}, 1 \le j \le i \le n$。故还要对误差协方差阵 Σ 作些假设，例如假设 Σ 已知或者 $\Sigma = \sigma^2 \Sigma_0$，而 Σ_0 已知等。

通常我们认为 n 次观测的误差是等方差的，且互不相关，即对误差协方差阵假设 $\Sigma = \sigma^2 I_n$(其中 I_n 表示 n 阶单位阵)，此假设条件称为 Gauss-Markov 条件，简写为 GM 条件。此时模型未知参数 $p+1$ 个，一般说来，只要观测值多于未知参数个数就可以对这些未知参数作出估计。

1.3.3 最小二乘法(least squared method)

考虑如下线性模型：

$$\begin{cases} Y = X\beta + e, \\ Ee = 0, Cov(e) = \sigma^2 I_n. \end{cases}$$

估计未知系数 β 的基本出发点是：参数的真值应该使模型误差 $e = Y - X\beta$ 达到最小。令 $Q(\beta) = \|e\|^2 = \|Y - X\beta\|^2$ 来度量模型误差的大小，则 β 的估计应最小化 $Q(\beta)$，即

$$\hat{\beta} = Arg \min_{\beta} \|Y - X\beta\|^2 。$$

上述确定未知参数 β 的方法称为最小二乘法，所得到的估计称为最小二乘估计(least squared estimate)，简写成 LSE。

例 1：回忆前面提到的曲线拟合问题，有 n 个点 $(x_i, y_i), 1 \le i \le n$，假设拟合的曲线形式已知为 $f(x, \beta)$(注意 β 是向量)，按照最小二乘法的原理未知参数的真值应该使的被拟合的曲线的纵向偏差的平方最小，即

$$\hat{\beta} = Arg \min_{\beta} \sum_{i=1}^{n} (y_i - f(x_i, \beta))^2 。$$

1.4 方差分析

在实际中经常要对在不同条件下进行试验或观察得到的数据进行分析，以判断不同条件对结果(响应变量)有无影响。此时协变量往往表示某种效应(条件)存在(成立)与否，因而往往取 0，1 两个值。在统计学上称这类分析为方差分析(analysis of variance，ANOVA)。

例 1：某农业科研机构欲比较三种小麦品种的优劣，设计了一种比较试验。为保证试验结果的客观性，他们选择了六块面积相等、土质肥沃程度一样的田地。每一种小麦播种在两块田地上，并给予几乎完全相同的田间管理。设用y_{ij}表示第i种小麦的第j块田地的产量，则可以对其作如下的分解：

$$y_{ij}=\mu+\alpha_i+e_{ij},\ i=1,2,3;\ j=1,2$$

其中μ表示总平均，α_i表示采用第i个小麦品种对产量的影响效应，e_{ij}表示其它为控制因素及各种随机误差的效应。

写成矩阵形式：

$$\begin{pmatrix}y_{11}\\y_{12}\\y_{21}\\y_{22}\\y_{31}\\y_{32}\end{pmatrix}=\begin{pmatrix}1&1&0&0\\1&1&0&0\\1&0&1&0\\1&0&1&0\\1&0&0&1\\1&0&0&1\end{pmatrix}\begin{pmatrix}\mu\\\alpha_1\\\alpha_2\\\alpha_3\end{pmatrix}+\begin{pmatrix}e_{11}\\e_{12}\\e_{21}\\e_{22}\\e_{31}\\e_{32}\end{pmatrix}。$$

若记 $Y=(y_{11}\quad y_{12}\quad y_{21}\quad y_{22}\quad y_{31}\quad y_{32})'$，

$\beta=(\mu\quad \alpha_1\quad \alpha_2\quad \alpha_3)'$，

$e=(e_{11}\quad e_{12}\quad e_{21}\quad e_{22}\quad e_{31}\quad e_{32})'$，

$$X=\begin{pmatrix}1&1&0&0\\1&1&0&0\\1&0&1&0\\1&0&1&0\\1&0&0&1\\1&0&0&1\end{pmatrix},$$

则可以写成：

$$Y=X\beta+e,$$

回到上节线性回归模型的形式，不过设计矩阵元素有其特点，非 0 即 1。例子所引进的模型是方差分析模型中最简单的一种，称为单因素方差分析，因为只涉及到“小麦品种”这一个因素的影响。在实际中还涉及到两个或多个因素的影响。

例 2：在例 1 中，一般情况下很难找到肥沃程度完全一样的田地。考虑到土质对产量的影响也是不可忽略的。一般，从若干试验田地中选取土质肥沃均匀的b块地(在试验设计中把这种块称为区组,block)，再将每块等分成 3 小块，称为试验单元，在每个试验单元上种植一种小麦。用y_{ij}表示第i种小麦的第j个区组田地的产量，则可以对其作如下的分解：

$$y_{ij}=\mu+\alpha_i+\beta_j+e_{ij},\ i=1,2,3;\ j=1,2,\cdots b$$

其中μ，α_i，e_{ij}的含义与例 1 一样，这里β_j表示第j个区组田地对产量的影响。

这样就得到一个两向分类模型。此模型也可以写成线性回归模型的形式：$Y=X\beta+e$，其中Y,e仿照前面例 1 的类似记号，$\beta=(\mu\quad \alpha_1\quad \alpha_2\quad \alpha_3\quad \beta_1\quad \cdots\quad \beta_b)'$，设计矩阵$X$为$3b\times(4+b)$矩阵，定义如下

$$X=\begin{pmatrix}1&1&0&0&1&&\\\vdots&\vdots&\vdots&\vdots&&\ddots&\\1&1&0&0&&&1\\1&0&1&0&1&&\\\vdots&\vdots&\vdots&\vdots&&\ddots&\\1&0&1&0&&&1\\1&0&0&1&1&&\\\vdots&\vdots&\vdots&\vdots&&\ddots&\\1&0&0&1&&&1\end{pmatrix}$$

方差分析模型认为由于各种因素的影响，研究所得的数据呈现波动状。造成波动的原因可分成两类，一是不可控的随机因素，另一是研究中施加的对结果形成影响的可控因素，例如例 1 中的“小麦品种”。在方差分析中，这些因素也称为因子(factor)，其影响称为效应(effect)，因子的不同状态称为水平(level)。

方差分析的基本思想是：通过分析研究不同来源的变异(方差)对总变异(方差)的贡献大小，从而确定可控因素对研究结果影响力的大小。其目的是通过数据分析找出对该事物结果有显著影响的因素，各因素之间的交互作用，显著影响因素的最佳水平等。

1.5 协方差分析

当知道有些协变量会影响响应变量，却不能够控制或不感兴趣时，可以在实验处理前予以观测，然后在排除这些协变量对观测变量影响的条件下，分析可控变量因素对观测变量的作用，从而更加准确地对控制因素进行评价。此方法称为协方差分析(analysis of covariance，ANCOVA)。

例 1：为研究三种饲料对猪的催肥效果，用每种饲料喂 8 头猪一段时间，测得每头猪的初始重量和增重(如下表)，现分析三种饲料对猪的催肥效果是否相同。

由于饲料是可以控制的，但猪的初始重量是人为无法控制的，因此要把猪的初始重量作为协变量考虑进来，进行协方差分析。令y_{ij}表示第i种饲料后第j头猪增重的体重，x_{ij}表示其初始体重，μ表示总平均，α_i表示采用第i种饲料对猪催肥的影响效应，e_{ij}表示其它为控制因素及各种随机误差的效应。

	初始体重	增重	饲料类型
1	15	85	1
2	13	83	1
3	11	65	1
4	12	76	1
5	12	80	1
6	16	91	1
7	14	84	1
8	17	90	1
9	17	97	2
10	16	90	2
11	18	100	2
12	18	95	2
13	21	103	2
14	22	106	2
15	19	99	2
16	18	94	2
17	22	89	3
18	24	91	3
19	20	83	3
20	23	95	3
21	25	100	3
22	27	102	3
23	30	105	3
24	32	110	3

可以建立如下模型：

$$y_{ij}=\mu+\gamma x_{ij}+\alpha_i+e_{ij},\ i=1,2,3;\ j=1,2\cdots 8。$$

令

$$Y=\left(y_{11}\ \cdots\ y_{18}\ \ y_{21}\ \cdots\ y_{28}\ \ y_{31}\ \cdots\ y_{38}\right)',$$

$$\beta=\left(\mu\ \ \gamma\ \ \alpha_1\ \ \alpha_2\ \ \alpha_3\right)',$$

$$e=\left(e_{11}\ \cdots\ e_{18}\ \ e_{21}\ \cdots\ e_{28}\ \ e_{31}\ \cdots\ e_{38}\right)',$$

则写成矩阵形式为：$Y=X\beta+e$，回到线性回归模型的形式，其中设计矩阵X为24×5的矩阵，定义为

$$X=\begin{pmatrix}1 & x_{11} & 1 & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 1 & x_{18} & 1 & 0 & 0\\ 1 & x_{21} & 0 & 1 & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 1 & x_{28} & 0 & 1 & 0\\ 1 & x_{31} & 0 & 0 & 1\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 1 & x_{38} & 0 & 0 & 1\end{pmatrix}$$

1.6 线性混合效应模型

为引入此模型，先从一个例子着手。

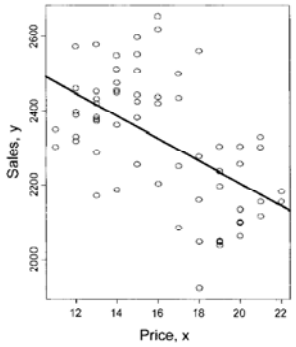
例 1：(销售量-价格模型)

为研究销售量与价格之间关系，选择K种商品，收集到这些商品的销售量与价格的数据$(y_i,x_i)_{i=1}^n$。按照传统线性回归模型，认为商品之间是没有差异的，n次观测是独立的，建立线性回归模型：

$$y_i=\alpha+\beta x_i+e_i,\ 1\le i\le n$$

这里假设e_i $i.i.d$为零均值，共同方差为σ_e^2。

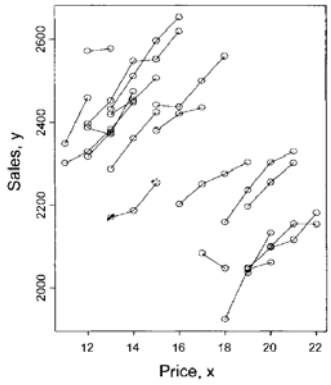
按照此模型，拟合回归直线，其斜率为负。



但事实上不同商品之间存在很大差异。如果按照商品种类将数据分组，记为(y_{ij}, x_{ij})，$i=1,2,\cdots K$，$j=1,2,\cdots n_i$，其中$\sum_{i=1}^{K} n_i = n$。对每组中的商品，如果分别建立其销售量与价格的线性回归模型，即假定每组商品有自己特定的模型

$y_{ij} = \alpha_i + \beta_i x_{ij} + e_{ij}$，$i=1,2,\cdots K$，$j=1,2,\cdots n_i$，

这里假设e_{ij} *i.i.d* 为零均值，共同方差为σ_e^2。将同组数据点连接的散点图(如下)，其显示大都斜率是正的。



相当于分组去拟合，得到不同组的商品各自的销售量与价格之间的关系。这似乎与最初的目标不一致。最初目标是想宏观的来看商品销售量与价格的关系，而不是特定某类商品销售量与价格的关系。

这样，如果假定每组中的价格对销售量的影响是一样的，不同在于各组截距不一样，得到如下模型：

$y_{ij} = \alpha_i + \beta x_{ij} + e_{ij}$，$i=1,2,\cdots K$，$j=1,2,\cdots n_i$，

这里假设e_{ij} *i.i.d* 为零均值，共同方差为σ_e^2。这样做法并没有解决上面所说的问题，不同的截距项$\{\alpha_i\}_{i=1}^{K}$也只反应了这K组不同商品之间的差异。并不代表总体商品之间的差异。注意如果$\alpha_i \equiv \alpha$，又回到最初的线性回归模型。

一种观点是：将这K组商品看成总体商品中随机抽取的K组商品，$\{\alpha_i\}_{i=1}^{K}$代表每组商品之间的差异。

为使模型简单，假设$\{\alpha_i\}_{i=1}^{K}$ *i.i.d* 均值为α，方差为σ_α^2。实际中，往往令$\alpha_i = \alpha + b_i$，此时$\{b_i\}_{i=1}^{K}$ *i.i.d* 均值为 0，方差为$\sigma_b^2 = \sigma_\alpha^2$。这样，我们就得到模型：

$y_{ij} = \alpha + \beta x_{ij} + b_i + e_{ij}$，$i=1,2,\cdots K$，$j=1,2,\cdots n_i$，

这里$\{b_i\}$ *i.i.d* 零均值，方差为σ_b^2，$\{e_{ij}\}$ *i.i.d* 为零均值，共同方差为σ_e^2且$\{b_i\}$与$\{e_{ij}\}$相互独立。

此模型中未知的α, β非随机，称为固定效应(fixed effect)，$\{b_i\}$未观察到，为随机效应(random effect)。当然，随机误差$\{e_{ij}\}$也可以称为随机效应。该模型称为线性混合效应模型(linear mixed models, LMM)。

另一种观点是：现在回过头来看我们的分组数据(y_{ij}, x_{ij})，$i=1,2,\cdots K$，$j=1,2,\cdots n_i$，其中$\sum_{i=1}^{K} n_i = n$。如果假定这n次观测是独立的，就回到前面建立的线性回归模型：$y_{ij} = \alpha + \beta x_{ij} + r_{ij}$，$i=1,2,\cdots K$，$j=1,2,\cdots n_i$，这里$r_{ij}$类似前面的*i.i.d*零均值方差为$\sigma_r^2$随机误差。

在实际中，往往会发现随机误差的变异(方差)σ_r^2所占比重过大。事实上，对不同组中的商品，假定观测独立是比较合理的。但对同一组商品，例如第i组商品，其$1,2,\cdots n_i$次观测应该是非独立的，具有某种相关性。因此固定i，误差$\{r_{ij}\}_{j=1}^{n_i}$独立的假定不是很合理。

为刻画其相关性，且使得模型简单，假设

$$r_{ij}=b_i+e_{ij},$$

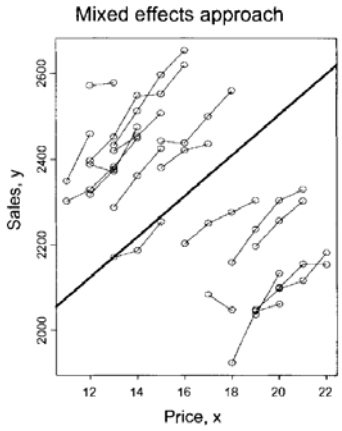
其中，b_i与e_{ij}独立都为零均值的随机变量，方差分别为σ_b^2和σ_e^2。从而$\sigma_r^2=\sigma_b^2+\sigma_e^2$，这也解释了为何以$r_{ij}$作为随机误差，其变异比重过大。组内相关系数

$$\rho=\frac{\sigma_b^2}{\sigma_b^2+\sigma_e^2}。$$

这样，我们也得到线性混合效应模型：

$y_{ij}=\alpha+\beta x_{ij}+b_i+e_{ij}$，$i=1,2,\cdots K$，$j=1,2,\cdots n_i$。

结果如下图：



此时$\rho=0.99$，表明变异主要是组内的变异。这也解释了为何两种方法估计的斜率符号相反。

再回过头来看我们的分组数据(y_{ij},x_{ij})，$i=1,2,\cdots K$，$j=1,2,\cdots n_i$，其中$\sum_{i=1}^{K}n_i=n$。对不同组中的商品，假定是独立的。但对同一组商品，例如第i组商品，其每次观测是相关。这样的数据我们成为纵项数据(longitudinal data)，在医学统计中也称为面板数据(panel data)。

一般线性混合效应模型形式为

$$Y=X\beta+Zb+e,$$

其中Y为$n\times1$观测向量，$X_{n\times p}$为已知设计矩阵，$\beta_{p\times1}$未知为固定效应，$Z_{n\times q}$为已知设计矩阵，$b_{q\times1}$为随机效应，且设$Eb=0$，$Cov(b)=D$非负定，e为随机误差且与b独立，$Ee=0$，$Cov(e)=R$为正定矩阵。

这样我们得到

$$Cov(Y)=ZDZ'+R。$$

对D，R的不同假设就可以得到不同的线性混合效应模型。