大数据丛书系列之五

总主编◎曾　羽　龙奋杰

大数据经典算法简介

DASHUJU JINGDIAN SUANFA JIANJIE

主　编◎胡文生　杨剑锋　张　豹

电子科技大学出版社

大数据丛书系列之五

总主编◎曾　羽　龙奋杰

大数据经典算法简介

DASHUJU
JINGDIAN SUANFA
JIANJIE

主　编◎胡文生　[illegible]刘锋　[illegible]代　豹

电子科技大学出版社

图书在版编目(CIP)数据

大数据经典算法简介 / 胡文生，杨剑锋，张豹主编.
-- 成都：电子科技大学出版社，2017.7
ISBN 978-7-5647-4815-9

Ⅰ. ①大… Ⅱ. ①胡… ②杨… ③张… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆CIP数据核字（2017）第175987号

大数据经典算法简介

胡文生　杨剑锋　张　豹　主编

策划编辑　杨仪玮　李燕芩
责任编辑　熊晶晶

出版发行　电子科技大学出版社
　　成都市一环路东一段159号电子信息产业大厦　邮编 610051
主　　页　www.uestcp.com.cn
服务电话　028-83203399
邮购电话　028-83201495

印　　刷　成都市火炬印务有限公司
成品尺寸　165mm × 240mm
印　　张　9.25
字　　数　162千字
版　　次　2017年7月第一版
印　　次　2017年7月第一次印刷
书　　号　ISBN 978-7-5647-4815-9
定　　价　35.00元

目　录

第一章　回归分析

事物的发生总是伴随着各种因果关系，在生产过程和科研实验中，我们发现某种现象或某种结果总是或多或少的与某个或某些因素相关联，但这种关联性又无法精确地使用数学模型来描述，只是能从数据上看出有这种趋势。对于具有这种特征的变量间的相关关系我们可以使用回归分析来研究分析。

回归分析(regression analysis)是一种用以处理两种或两种以上的变量之间的相关关系的一种很有效的统计分析方法[1]。回归分析按考察因素不同有不同的分类，其中比较常见的分类方法有：回归分析按所涉及自变量的多少，可分为一元回归分析和多元回归分析；按自变量与因变量之间的关系类型，可分为线性回归分析和非线性回归分析。

需要指出的是，所有的回归分析模型都可以理解为两个部分[2]，即

观测值＝结构项＋随机项

其中，观测值表示因变量的实际值；结构项表示因变量与自变量间的结构关系，表现为预测值；随机项表示观测值中未被结构项所解释而剩余的部分。

1.1　一元线性回归分析

1.1.1　一元线性回归模型

如果随机变量 Y 和变量 X 之间服从如下关系

$$Y=a+bX+\varepsilon \tag{1-1}$$

对已有的 (X,Y) 的 n 对实验数据 (x_i,y_i)，$i=1,2,\ldots,n$，满足

$$y_i=a+bx_i+\varepsilon_i \tag{1-2}$$

其中，a,b 是未知参数，ε_i 是随机误差项，$\varepsilon_i\sim N(0,\sigma^2)$，$i=1,2,\ldots,n$。我们称变量 Y 和变量 X 服从一元线性回归模型(又称一阶模型)[3]。

1　高惠璇. 统计计算[M]. 北京：北京大学出版社，1995(第 6 章).

2　谢宇. 回归分析：修订版[M]. 北京：社会科学文献出版社，2006(第 3 章).

3　何正风. MATLAB 概率论与数理统计分析：第 2 版[M]. 北京：机械工业出版社，2012(第 6 章).

对于上述一元线性回归模型，在实际应用中一般主要考虑以下问题：在已知的观测值(x_i,y_i)，$i=1,2,\ldots,n$的基础上，对未知参数a,b,σ^2进行估计，并对a,b的估计值进行检验，进而预报变量Y的值。

1.1.2 参数估计

未知参数a,b的最小二乘估计

对于观测数据(x_i,y_i)，$i=1,2,\ldots,n$，作残差平方和：$Q(a,b)=\sum_{i=1}^{n}(y_i-a-bx_i)^2$。我们选择使得$Q(a,b)$取得最小值的$(\hat{a},\hat{b})$作为未知参数$(a,b)$的最小二乘估计，为此，分别求$Q(a,b)$对$a$和$b$的一阶偏导数并令其为零，得

$$\begin{cases}\dfrac{\partial Q}{\partial a}=-2\sum\limits_{i=1}^{n}(y_i-a-bx_i)=0\\[2ex]\dfrac{\partial Q}{\partial b}=-2\sum\limits_{i=1}^{n}(y_i-a-bx_i)x_i=0\end{cases}\tag{1-3}$$

整理得

$$\begin{cases}na+b\sum\limits_{i=1}^{n}x_i=\sum\limits_{i=1}^{n}y_i\\[2ex]a\sum\limits_{i=1}^{n}x_i+b\sum\limits_{i=1}^{n}x_i^2=\sum\limits_{i=1}^{n}x_iy_i\end{cases}\tag{1-4}$$

我们记$\bar{x}=\dfrac{1}{n}\sum\limits_{i=1}^{n}x_i$，$\bar{y}=\dfrac{1}{n}\sum\limits_{i=1}^{n}y_i$，则方程组的解可表示为

$$\begin{cases}\hat{b}=\dfrac{\sum\limits_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}\\[2ex]\hat{a}=\bar{y}-\hat{b}\bar{x}\end{cases}\tag{1-6}$$

进而可得回归直线方程

$$\hat{Y}=\hat{a}+\hat{b}X\tag{1-7}$$

其中，$\hat{Y}$表示$\hat{a}$和$\hat{b}$确定之后对于给定的X而相对应的Y的预报值(predicted value)，又称Y的拟合值(fitted value)或回归值(regression value)。公式(1-7)称为预报方程(predicted equation)，又称拟合方程(fitted equation)或经验回归方程；与之对应的直线称为拟合直线(fitted straight line)，又称回归

直线(regression straight line)[4]。

将 $\hat{a}=\bar{y}-\hat{b}\bar{x}$ 代入回归直线方程,则

$$\hat{Y}-\bar{y}=\hat{b}(X-\bar{x}) \tag{1-8}$$

除了上述的 a,b 最小二乘估计之外,我们还可以使用最小二乘估计的矩阵算法和极大似然估计等方法对未知参数 a,b 进行估计,在此不再一一列出。

σ^2 的估计

由于 $\sigma^2=D(\varepsilon)=E(\varepsilon^2)$,故可用 $\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2$ 对 σ^2 进行估计,代入 (a,b) 的估计量 $(\hat{a},\hat{b})$,可得

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{a}-\hat{b}x_i)^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2-\hat{b}\,\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{1-9}$$

1.1.3 案例分析

从上述内容中可知,根据样本中自变量 X 的值,可以对回归直线上相应的因变量 Y 的值进行预报。下面以一个例子来解释上述过程。

某种合金强度与碳含量有关,我们在生产试验中收集了该合金的强度 Y 与碳含量 X 的数据,试建立 Y 与 X 的函数回归模型。

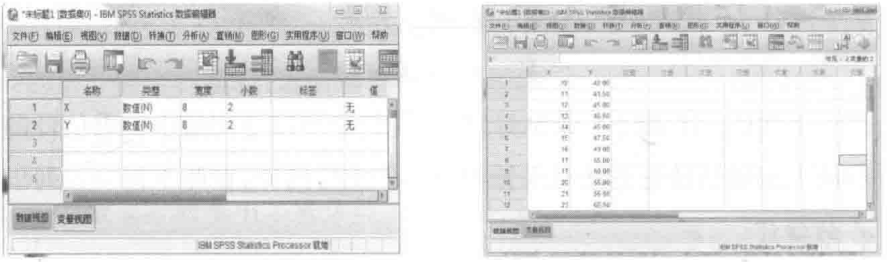
表 1-1 某种合金强度与碳含量数据

碳含量 X	合金的强度 Y(MPa)	碳含量 X	合金的强度 Y(MPa)
0.10%	42.0	0.16%	49.0
0.11%	41.5	0.17%	55.0
0.12%	45.0	0.18%	50.0
0.13%	45.5	0.20%	55.0
0.14%	45.0	0.21%	55.5
0.15%	47.5	0.23%	60.5

下面我们首先给出 SPSS 软件的操作方法:打开 SPASS 软件,选择变量视图(Variable View)[如图 1-1(a)所示],添加自变量 X 和因变量 Y,再选择数

4 钱俊龙. 概率论与应用统计[M]. 北京:中国统计出版社,1992(第 8 章).

据视图(Data View),输入数据[如图 1-1(b)所示]。



(a) (b)

图 1-1 SPSS 主界面

点击分析(Analyze)→回归(Regression)→线性回归(Linear Regression)[如图 1-2(a)所示],进入线性回归分析界面,在因变量栏(Dependent)选择 Y,自变量栏(Independent)选择 X[如图 1-2(b)所示],点击确定。

(a) (b)

图 1-2 分析操作及变量选择

在统计量(Statistics)中选择相应的统计量(如系数估计、系数置信区间、模型拟合、描述性等),点击继续→确定,即可输出相关数据(如表 1-2 所示)。

表 1-2 输出结果

模型	非标准化系数		标准系数	t	Sig	B 的 95.0%置信区间	
	B	标准误差	试用版			下限	上限
1 (常量)	26.790	1.947	966	13.761	.000	22.452	31.128
X	142.867	11.999		11.907	.000	116.132	169.602

由表 1-2 可知,回归方程为 $Y=26.79+142.87X$,相关系数为 0.716。

回归直线如图 1－3 所示。

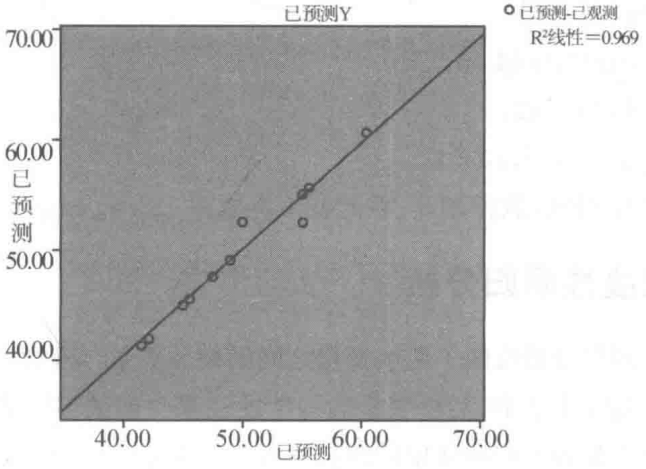


图 1－3　回归直线

当然，我们也可以在 MATLAB 软件中调用函数 regress 求解。

调用格式：

b=regress(y,x)：输出线性回归方程的系数的估计值向量 b。

[b,bint]=regress(y,x)：附加输出系数估计值的置信度为 95%的置信区间 bint。

[b,bint,r,rint]=regress(y,x)：附加输出残差向量 r。

[b,bint,r,rint,stats]=regress(y,x)：附加输出向量 stats，包括判定系数 R^2、统计量 F 的观测值、检验的 p 值和方差估计值 σ2。

源程序代码：

```
clear all;
x1;
x2;
y;
x=[ones(12,1),x2];
figure;
plot(x2,y,'*');
[b,bint,r,rint,stats]=regress(y,x);
```

```
b,bint,stats,
figure(2);
rcoplot(r,rint);hold on;
z=b(1)+b(2)*x2;
plot(x2,y,'*',x2,z,'r');
```

输出结果与 SPSS 软件相同，在此就不再累述。

1.2 多元线性回归分析

一元线性回归分析反映了两个变量之间的相关关系，是回归分析的一种特殊形式。但在实际生活中，某种现象的发生往往是受到诸多因素影响的结果。下面介绍一种更常为人们所使用的回归分析——多元线性回归分析，它是一元线性回归分析的推广，是研究一个或多个变量（因变量）相对于其他多个变量（自变量）之间的相关关系的数学模型。在实际问题中，如果只考察一个变量（因变量）与其他多个变量（自变量）之间的相关关系，则称为多元回归问题（一对多型）；如果需同时考察多个变量（因变量）相对于其他多个变量（自变量）之间的相关关系，则称为多变量多元回归问题（多对多型）[5]。

鉴于实际应用中，一对多型的多元回归问题更加多见，所以本节只介绍一对多型的多元线性回归分析模型。

1.2.1 多元线性回归模型

由于多元线性回归分析是一元线性回归分析的推广，故其模型的建立、参数的估计等方面的思路方法和一元线性回归分析都是一致的。

多元线性回归分析研究的是因变量 Y 与 n 个自变量 $X_1, X_2, \ldots, X_n$ 之间的相关关系，一般假设因变量 Y 为随机变量，自变量 $X_1, X_2, \ldots, X_n$ 为确定变量，因此可建立多元线性回归模型

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \ldots + a_n X_n + \varepsilon \tag{1-10}$$

其中，$a_i, i = 0, 1, \ldots, n$ 是未知参数，称为总体参数；ε 是随机误差项，代表自变量以外被忽略的或无法考虑的其他随机影响因素，一般对预报值不会产生系统性的偏差效应。

5 高惠璇. 统计计算[M]. 北京：北京大学出版社，1995(第 6 章).

对于一个实际问题的 m 组样本数据

$$\begin{array}{cccc} Y\backslash X & x_1 & x_2 & \cdots x_n \\ y_1 & x_{11} & x_{12} & \cdots x_{1n} \\ y_2 & x_{21} & x_{22} & \cdots x_{2n} \\ \vdots & \vdots & \vdots & \\ y_m & x_{m1} & x_{m2} & \cdots x_{mn} \end{array} \tag{1-11}$$

则线性回归模型可表示为

$$\begin{cases} y_1 = a_0 + a_1 x_{11} + a_2 x_{12} + \ldots + a_n x_{1n} + \varepsilon_1 \\ y_2 = a_0 + a_1 x_{21} + a_2 x_{22} + \ldots + a_n x_{2n} + \varepsilon_2 \\ \cdots\cdots \\ y_m = a_0 + a_1 x_{m1} + a_2 x_{m2} + \ldots + a_n x_{mn} + \varepsilon_m \end{cases} \tag{1-12}$$

记矩阵

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1n} \\ 1 & x_{21} & \cdots & x_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{m1} & \cdots & x_{mn} \end{bmatrix} \quad a = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{pmatrix}$$

则(1-12)可写成矩阵形式

$$y = Xa + \varepsilon \tag{1-13}$$

其中，y 为因变量向量，X 为所有自变量和一列常数所组成的矩阵[$\mathrm{rank}(X^T) = n$]，a 为总体参数向量，ε 为随机误差项向量(这里 $\varepsilon_i \sim N(0, \sigma^2)$，$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0, (i = 1, \ldots, m, i \neq j, m > n)$。

1.2.2 参数估计

总体参数 $a_0, a_1, \ldots, a_n$ 的最小二乘法

现实问题中，多元线性回归模型是根据已有理论甚至是经验总结而“假设设定”的，其总体参数 $a_i (i = 0, 1, \ldots, n)$ 自然是未知的，我们只能基于具体样本的观测数据来构建合适的样本的回归模型，进而推断总体的回归模型，基于上述给出的 m 组样本数据，估计总体参数。与一元线性回归分析一样，采用最小二乘估计，计算残差平方和

$$Q(a) = (y - Xa)^T (y - Xa) = \sum_{j=1}^{m} (y_j - a_0 - \sum_{i=1}^{n} a_i x_{ji})^2 \tag{1-14}$$

我们仍使用最小二乘法计算向量 a 的估计量 $\hat{a}$，即向量 $\hat{a}$ 满足

$$Q(\hat{a}) = \min_{a} Q(a) \tag{1-15}$$

为此，我们求偏导数 $\dfrac{\partial Q(a)}{\partial a_k} = 2 \sum_{j=1}^{m} (y_j - a_0 - \sum_{i=1}^{n} a_i x_{ji}) x_{jk}, k = 0, 1, \ldots, n$，并

令其为零，得

$$\sum_{j=1}^{m}(y_j-a_0-\sum_{i=1}^{n}a_i x_{ji})x_{jk}=0,k=0,1,\ldots,n$$

整理可得方程组

$$a_0\sum_{j=1}^{m}x_{jk}+\sum_{i=1}^{n}a_i\sum_{j=1}^{m}x_{ji}x_{jk}=\sum_{j=1}^{m}x_{jk}y_j,k=0,1,\ldots,n \tag{1-16}$$

写成矩阵形式为

$$X^TXa=X^Ty \tag{1-17}$$

因为 $\mathrm{rank}(X^TX)=\mathrm{rank}(X^T)=n+1$，故$(X^TX)^{-1}$存在，则

$$\hat{a}=(X^TX)^{-1}X^Ty \tag{1-18}$$

$\hat{a}=(\hat{a}_0,\hat{a}_1,\ldots,\hat{a}_n)$即为总体参数的最小二乘估计，满足：线形性、无偏性、最小残差性和正态性[1]。

误差方差 σ^2 的估计

将自变量的观测值代入回归方程，即得因变量的预报值为

$$\hat{y}\overset{\Delta}{=}(\hat{y}_1,\hat{y},\ldots,\hat{y}_m)=X\hat{a}$$

我们称

$$e\overset{\Delta}{=}y-\hat{y}=y-X\hat{a}=[E-X(X^TX)^{-1}X^T]y=[E-H]y \tag{1-19}$$

为残差向量（又称剩余向量），其中 E 为 m 阶单位矩阵，$H=X(X^TX)^{-1}X^T$ 为 m 阶幂等矩阵。称

$Q(e)=e^Te=(y-X\hat{a})^T(y-X\hat{a})=y^T[E-H]y=y^Ty-\hat{a}^TX^Ty$

为残差平方和。

由于 $E(y)=Xa$ 且$[E-H]y=0$，则

$Q(e)=e^Te=(y-X\hat{a})^T[E-H](y-X\hat{a})=\varepsilon^T[E-H]\varepsilon$

于是

$$\begin{aligned}E(e^Te)&=E(tr(\varepsilon^T[E-H]\varepsilon))=tr(([E-H])E(\varepsilon^T\varepsilon))\\&=\sigma^2tr(E-X(X^TX)^{-1}X^T)\\&=\sigma^2(m-tr(X(X^TX)^{-1}X^T))\\&=\sigma^2(m-n)\end{aligned}$$

其中，$tr(X)$表示矩阵 X 的迹。因此

$$\hat{\sigma}^2\overset{\wedge}{=}\frac{1}{m-n}e^Te \tag{1-20}$$

为 σ^2 的一个无偏估计。

1.2.3 案例分析

下面以某气象台为预报该地某月平均气温 Y，选择了与之密切相关的四个气象要素(X_1,X_2,X_3,X_4)作为预报因子（自变量），以历史上 22 年的数据作为样本，原始数据如表 1－3 所示，利用多元回归分析的方法求月平均气温的预报

方程，并给出当 $X_1=0, X_2=-9, X_3=9, X_4=34$ 时的当月平均气温的预报值。

表 1-3　某地过去 22 年的月平均气温及相关因素的数据样本

样本号 / 变量	X_1	X_2	X_3	X_4	Y
1	4	21	1	26	23.9
2	4	12	0	31	24.6
3	0	10	7	37	22.4
4	0	−25	6	28	20.8
5	7	9	6	30	21.9
6	4	12	5	33	22.5
7	4	5	5	33	23.6
8	2	19	7	27	23.1
9	0	17	4	34	23.0
10	0	9	4	35	23.2
11	2	2	8	29	24.7
12	0	−2	9	34	22.0
13	8	−4	4	36	21.1
14	1	−35	2	29	23.0
15	0	−35	5	29	23.1
16	0	8	4	36	24.1
17	1	10	4	34	22.6
18	0	−11	4	28	22.5
19	0	−6	5	37	21.4
20	2	11	1	35	21.2
21	0	−33	5	29	22.4
22	1	−3	1	29	25.1
均值	1.818 2	−0.409 1	4.181 8	31.772 7	22.827 3

输入变量和观测值，点击分析(Analyze)→回归(Regression)→线性回归(Linear Regression)，进入线性回归界面，选择变量(如图 1-4 所示)，在统计量中选择统计量(如系数估计、系数置信区间、模型拟合、描述性、相关系数等)，点击继续(Continue)→确定(Ok)，即可输出相关结果(如表 1-4 所示)。

图 1－4　变量选择

表 1－4　部分结果

模型	非标准化系数		标准系数	t	Sig	B 的 95.0%置信区间		相关性			共强性统计量	
	B	标准误差	试用版			下限	上限	零阶	偏	部分	容量	VIF
1(常量)	28.124	2.347		11.983	000	23.172	33.076					
			−260					−666	−280	241	−858	1.165
X_1	−129	108		−1.202	246	−357	098					
			416					231	410	372	800	1.260
X_2	028	−015		1.853	081	−0.04	−061					
			−257					−287	−291	−252	961	1.041
X_3	−130	103		−1.256	−226	−348	−088					
			−415					−427	−427	−391	987	1.128
X_4	141	075		−1.948	068	−293	−012					

由输出结果可知，回归方程为

$$Y=28.124-0.129X_1+0.28X_2-0.13X_3-0.141X_4$$

相关系数为 0.562。检验输出结果如表 1－5 所示，偏回归平方和为 9.308，残差平方和为 20.176，离差平方和为 29.484。

表 1－5　Anova[b]

模型	平方和	df	均方	F	Sig
1　回归	9.308	4	2.327	1.961	147[2]
减差	20.176	17	1.187		
总计	29.484	21			

a. 预测变量(常量)：X_4，X_1，X_3，X_2。

b. 因变量：Y。