浅析最小二乘法及其在数学建模中的应用

□梁博尧　华南师范大学附属中学

【摘要】　在利用数学建模方法解决现实生活中的问题时，常常需要将问题中的数据离散化，通过不同的数学分析方法建立合适的函数关系。最小二乘法在利用观测数据解决这类函数方程的问题中有重要的应用，它通过误差平方和最小的限定条件可以求得最佳拟合曲线，并对其误差进一步分析。本文通过对最小二乘法的原理进行分析，并阐述了这种方法在数学建模中的应用，利用一元线性回归分析方法结合多组观测数据进行数值实验，最后简单概括最小二乘法的重要地位。

【关键词】　最小二乘法　数学建模　一元线性回归　matlab

一、数学建模与最小二乘法

随着计算机科学的不断发展，各类数学方法不仅在物理，化学等自然科学中有很大的应用，同时在经济，管理等社会科学起到重要的作用。所谓数学技术已经从最初的基础数学逐渐成为当代高新技术的重要组成部分。在众多数学分析方法中，数学建模往往可以和我们实际问题紧密的结合。一般来讲，数学建模并不是现实问题的直接翻版，我们通常需要对问题进行更深入，更细致的观察和分析，与此同时又需要灵活运用各种数学知识。在整个过程中，我们首先需要构建系统模型结构，很多时候都是从已知模型入手再进行实验。解模型的过程就是利用各种数学知识和技巧，从已知系统模型结构中建立相应的函数关系，通过利用系统的输入和输出数据来估计模型参数。对模型参数的估计往往离不开数据与曲线拟合，以及实验误差分析，这里的重点便是最小二乘法的运用。

最小二乘法最早是由数学家勒让德提出，最初应用于求解天文学方面的问题，比如确定行星的天体轨迹等。勒让德在其著作《计算慧星轨道的新方法》中对最小二乘法的优点进行了阐述，但并没有进行相应的误差分析。随着最小二乘法逐渐在不同领域应用，辛普森，拉普拉斯等人对该方法进一步修正，最后“数学王子”高斯将其与正太分布相结合，具体阐明其误差分析手段，成为数理统计史上最重要的成就之一，正如美国统计学家斯蒂格勒所说，“最小二乘法之于数理统计学犹如微积分之于数学”。

最小二乘法从本质上来讲是一种近似拟合的方法，可以对实际事件通过进行大量的观测获得最佳估计或者说最可能的结果。

比如我们常见的线性方程 y=a+bx, 其中自变量 x 与因变量 y 呈线性关系，通过对满足这种关系的实际事件其进行 n (n >2) 次观测，我们可以获得 n 组数据，实际中观测数据并不能够完全拟合线性方程，所以如果把 n 对数据代入方程来求解参数 a 与 b 的值无法得到确定解。最小二乘法为解决这类问题提供了一种求解方法，基本思想就是寻找最接近观测点的直线，它通过寻求误差平方加和最小并确定数据的最佳函数匹配。利用最小二乘法可以简便地求得满足观测数据的最佳函数关系，并对之加以利用来求解未知的数据，尽可能的使这些求得的数据与未观测到的实际数据之间误差的平方和最小。

总的来说最小二乘法不仅仅是 19 世纪最重要的统计分析方法，相关回归分析、方差分析等数理统计方法都以最小二乘法为理论基础。

二、最小二乘法的原理

假设自变量 x 和因变量 y 是具有某种相关关系的物理量，这样它们之间的函数关系可用下式表示：

$y=f(x,c_1,c_2\cdots\cdots c_N)$

上式中 $c_1,c_2,\cdots\cdots c_N$ 为 N 个待定常数，函数关系式形式已经确定但是具体关系并没有确定，即曲线形式确定而曲线形状未定。

为求解得到具体函数关系，我们可以同时测得 x 和 y 的数值，假设一共获得 m 组观测结果：$(x_1，y_1)(x_2，y_2)\cdots\cdots(x_m，y_m)$，根据这 m 组观测值可以确定常数 $c_1,c_2\cdots\cdots c_N$。假设 x 和 y 的函数关系最佳表达形式为：

$$\hat{y}=f(x,\hat{c_1},\hat{c_2}.....\hat{c_N})$$

上式中，$\hat{c_1},\hat{c_2}.....\hat{c_N}$ 是 $c_1,c_2\cdots\cdots c_N$ 的最佳估计值。如果在观测数据与理论结果之间不存在误差，那么 m 组观测值都应落在函数曲线式上，但往往在确定这类关系时观测数据并不能与理论曲线完全拟合而存在误差，则有：

$$e_i=y_i-\hat{y_i}$$

上式中 i=1,2……m，e_i 是残差，为误差的实测值。如果 m 对观测值中有比较多点 $(x_i，y_i)$ 落到曲线上，那么我们求解得到的曲线就可以较为准确地反映被测物理量 x 与 y 之间的关系，否则这种函数关系便不适用。当观测值 y 落在曲线上的概率达到最大时，且观测误差服从 $N(\sigma,\hat{y_i})$ 的正态分布，则 $P(e_1,e_2\cdots\cdots e_m)$ 概率为：

$$P(e_1,e_2.....e_m)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\sum_{i=1}^{m}\frac{(y_i-\hat{y_i})^2}{2\sigma^2}\right]$$

如果 $P(e_1,e_2\cdots\cdots e_m)$ 达到最大时，那么我们求解的曲线便为最佳形式。那么当下式达到即最小残差平方和最小时，我们可以得到最优解：

$$\delta=\sum_{i=1}^{m}(y_i-\hat{y_i})^2=\sum_{i=1}^{m}e_i^2$$

我们通过对上式求导，当导数为 0 时，便可得到最小值即：$\frac{\partial\delta}{\partial c_1}=0,\frac{\partial\delta}{\partial c_2}=0.....\frac{\partial\delta}{\partial c_N}=0$

将其展开可以得到如下表达：

$$\sum_{i=1}^{m}[(y_i - f(x_i, \hat{c}_1, \hat{c}_2 \ldots\ldots \hat{c}_N)](\frac{\partial f}{\partial c_1}) = 0$$
$$\sum_{i=1}^{m}[(y_i - f(x_i, \hat{c}_1, \hat{c}_2 \ldots\ldots \hat{c}_N)](\frac{\partial f}{\partial c_2}) = 0$$
……
$$\sum_{i=1}^{m}[(y_i - f(x_i, \hat{c}_1, \hat{c}_2 \ldots\ldots \hat{c}_N)](\frac{\partial f}{\partial c_N}) = 0$$

上式便为正规方程，通过求解该方程组可以得到待定常数，这种方法即为最小二乘法解。函数关系的不同，该方程组的具体表达形式不同，例如在求解一元线性回归拟合和多元线性回归拟合时，我们需要求解的待定参数的数量也不同。

三、最小二乘法应用

在利用数学建模解决实际问题时，我们往往将问题抽象成数学关系式。首先需要确定假设的条件，然后利用适当的数学方法来刻画不同变量间的关系，优先选择简单的数学工具。通过对问题分析和测试，我们能够获得相应的数据资料，对数学模型中的待定参数近似计算，然后再对结果进行分析。所以结合最小二乘法的原理，我们不难发现在求解模型以及将模型理论结果与实际情形进行比较的过程中，均可以利用最小二乘法验证和评价模型的准确性、合理性和适用性。

本文利用一元线性回归模型，从回归分析的角度建立线性模型，来讨论利用这种方法来解决数学建模问题中最小二乘法的运用。回归分析的主要目的是要通过样本的回归方程不断修正待定参数来尽可能准确的估计总体的数学关系。所以我们以简单的一元线性回归模型为例，来阐述运用最小二乘法进行模型的参数估计。其他分析方法如多元线性回归分析，非线性回归分析等均可以类推求解。

假定数学模型中仅有两个变量 x 和 y，且关系为线性形式：

$y=a+bx$

上式中 a，b 为需要确定的待定常数。这类方程即为一元线性回归方程，方程中的参数 a， b 称为线性回归系数。同时假定观测数据为 $(x_1, y_1)(x_2, y_2)\cdots\cdots(x_n, y_n)$，且实验结果与理论结果之间存在误差，将观测数据带入方程左右两边不等即观测点不能落在回归直线上。利用最小二乘法的原理，要获得最佳拟合曲线需要保证观测数据点与回归直线的偏差平方和最小，通过求解偏导数使其为 0 我们可以得到：

$$\frac{\partial\sum_{i=1}^{n} v_i^2}{\partial a} = -2\sum(y_i - a - bx_i)$$
$$\frac{\partial\sum_{i=1}^{n} v_i^2}{\partial b} = -2\sum(y_i - a - bx_i)\cdot x_i$$

上式中 v_i^2 为误差平方，进一步整理可以得到：

$$\sum_{i=1}^{n} y_i - na - b\sum_{i=1}^{n} x_i = 0$$
$$\sum_{i=1}^{n} y_i x_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 = 0$$

且误差平方和对 a 和 b 的二阶偏导为：

$$\frac{\partial^2\sum v_i^2}{\partial a^2}\cdot\frac{\partial^2\sum v_i^2}{\partial b^2} - (\frac{\partial^2\sum v_i^2}{\partial a\partial b})^2$$
$$= 4[n\sum x_i^2 - (\sum x_i^2)]$$
$$= 4n[\sum x_i^2 - (\sum x_i^2)/n]$$
$$= 4n\sum(x_i - \bar{x}_i)^2 > 0$$

所以求得待定常数 a 和 b 可使一阶偏导为极小值，继而得到的拟合曲线 $y=a+bx$ 就是利用最小二乘法得到的最佳曲线。同理我们也可以利用最小二乘法评价参数 a 和 b 的精度以及拟合曲线的相关性。

为验证理论推导的准确性，本文利用数学软件 matlab 对线性回归模型进行数值模拟，matlab 语言提供了一个函数，可以直接完成线性曲线拟合并绘制相应图即 polyfit 函数。模型输入量为 x、y、n，其中 x、y 即为需要建立数学关系变量，将观测值以数组的形式输入函数中，n 为多项式的次数，输出得到的是多项式系数的行向量。

表 1 为给出的观测数据点 (x_i, y_i)，利用 matlab 将观测数据绘制成散点图如图 1 所示。

xi	−2.51	−1.72	−1.13	−0.82	0	0.11	1.52	2.71	3.63
yi	−192.8	−85.51	−36.14	−26.54	−9.11	−8.42	−13.15	6.52	68.06

表 1　线性回归观测数据

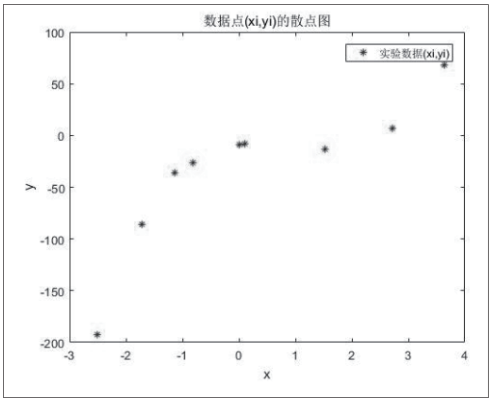

图 1 观测数据离散图

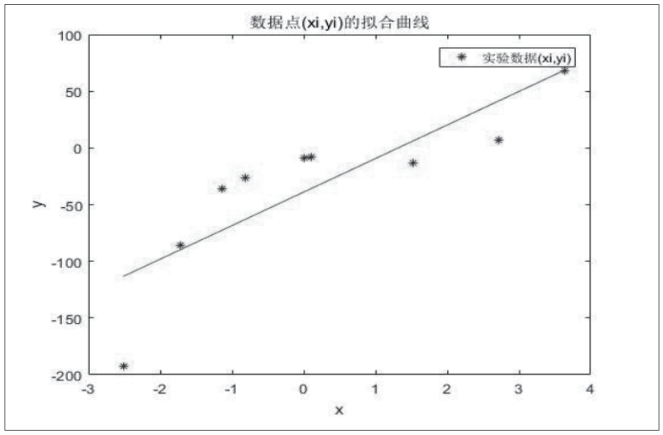

图 2 一元线性拟合曲线图

基于虚拟机技术的计算机网络实验教学探讨

□刘芹　山东商业职业技术学院　潘大勇　山东鲁商物流科技有限公司

【摘要】　在《计算机网络技术》课程的授课过程中，接近一半的课时是实验内容。而目前授课机房的实验环境，在支撑实验课程的开展方面还有一定的局限性。本文提出采用虚拟机的方法搭建实验环境，并且详细介绍了在新搭建的实验环境下，如何开展实验教学。
【关键字】　计算机网络技术　虚拟机　教学

一、前言

1.1 研究背景

互联网的飞速发展，使得众多领域享受到它所带来的便利、快捷和高效。而互联网迅速发展的基础是计算机网络技术的推广及应用。没有基础的网络设备、经典的各类协议、基本的组网规则，就谈不上互联网的迅猛发展，更没有我们现在的日新月异。在这样的大背景下，很多学科和专业都需要计算机网络技术知识体系的深入学习。在学习中，只有单纯的理论内容远远不够，实验内容的重要性也越来越突出。因此，作为授课教师，教学重点就放在了实验课的教学设计中。

1.2 实验教学现状

目前的实验教学授课所用的机房，学生机大多采用了还原卡功能，这些环境在常规教学中起到了很好的管理作用，例如图片及动画制作、数据库 SQL 语句的书写及运行、办公自动化软件的学习以及作业提交等。但这些学生机，都具备一个共同的特点——重启后原有系统中的作业会清除。而计算机网络实验教学中，均需要通过多次重启计算机来完成实验内容，这一问题便成了此类实验体系发展的瓶颈。

1.3 虚拟机技术简介

虚拟机（Virtual Machine）技术，在一定程度上，可以提供一种解决方法。虚拟机是可以安装在操作系统中的一个软件，该软件相当于一个容器，在这个容器体系内，可以安装多个不同版本的操作系统却不会影响原有操作系统。常见的虚拟机为 VM ware 这一软件，软件有多个版本。目前学生机安装的是 32 位的 Win7 操作系统。有些版本的虚拟机在 Win7 上运行时，需要进行相应的硬件配置方可使用，给实验带来很多不便，经过笔者的多次验证，最终选用了 VMware Workstation 10 精简版，该版本可以顺畅地安装在 32 位的 Win7 操作系统上。

二、《计算机网络技术》课程相关实验

计算机网络课程的实验内容有很多，其中最有代表性的两个实验为 DHCP 服务与 FTP 站点架设。

2.1 DHCP 服务

众所周知，在计算机网络的实际应用中，更多的是服务器网络而不是对等网络，因此，在网络中会根据网络功能的要求架设一些专用的网络服务器，从而为用户提供多种网络服务，其中最常见的服务就是 DHCP 服务，即动态主机配置协议（Dynamic Host Configuration Protocol），通过建立 DHCP 服务器，可以给网络中的计算机动态地分配地址以及在这个基础上进行一些网络权限等的高级设置。

2.2 FTP 站点架设

网络的功能之一是可以实现资源共享和文件的传输，在常规的局域网中，用户可以通过“网上邻居”访问网络中其他计算机上的共享资源。而在 Internet 中，因为很多因素的限制，往往不能通过共享文件夹的方式来实现文件的共享。这时，就用到了经典的 FTP 协议，即文件传输协议（File Transfer Protocol）。可以通过架设 FTP 站点的途径来实现文件的共享。

三、利用虚拟机开展实验教学

Y=29.57x–38.89

通过计算我们可以得到一元线性拟合曲线的表达形式为 y=29.57x–38.89，由图 2 我们可以看到拟合曲线与观测数据的变动趋势基本保持一致，但是数据点落在曲线上数量较少。对这个问题我们做以下分析，首先 x_i，y_i 相关系数为 0.84 并没有达到 1 即数据间不是完全正相关，所以观测数据点不能完全落在直线上。其次一元线性拟合结果精度较差，若采用非线性拟合效果会更好。

总而言之，最小二乘法可以尽可能满足各个近似条件并使其达到误差方差最小，通过上述数值实验，可以看出利用最小二乘法，我们可以很好的解决数学建模中利用观测数据建立函数方程的问题。但针对数据间不同的数学关系以及相关性等，为达到最好的拟合效果，我们需要采用的拟合方法也不尽相同。

参 考 文 献

[1] 陆健 . 最小二乘法及其应用 [J]. 中国西部科技报 ,2007.
[2] 魏宗舒 . 概率论与数理统计教程（第二版）(M). 北京：高等教育出版社，2010.
[3] 王萼芳 . 高等代数（第三版）(M). 北京：高等教育出版社，2009.
[4] 陈希孺 . 最小二乘法的历史回顾与现状 (J). 中国科学院研究生院学报 ,1998.
[5] 贾小勇 . 最小二乘法的创立及其思想方法 (J). 西北大学学报 (自然科学版),2006.