DOI:10.16661/j.cnki.1672-3791.2014.29.001

多元线性回归分析实例分析

王华丽
(湖北文理学院数学与计算机科学学院 湖北襄阳 441052)

摘 要:多元线性回归是简单线性回归的推广,研究的是一个变量与多个变量之间的依赖关系。作为质量统计软件领域的领导者,MINITAB是一个精确的、强大的、使用方便的统计软件。多元回归分析预测法,是指通过对两个或两个以上的自变量与一个因变量的相关分析,建立预测模型进行预测的方法。当自变量与因变量之间存在线性关系时,称为多元线性回归分析。该文通过一个具体实例介绍如何运用MINITAB软件,建立儿子身高与父母身高、年锻炼次数的多元线性回归模型,并对MINITAB的输出结果进行分析,得出方程效果良好的结论。

关键词:MINITAB软件 多元线性回归 显著性 实例分析
中图分类号:0212 文献标识码:A 文章编号:1672-3791(2014)10(b)-0022-02

回归分析是数据分析中使用很多的一种方法。回归分析是定量的给出变量间的变化规律,它不仅提供变量间的回归方程,而且可以判断所建立回归方程的有效性。在方程有效性的前提下,可以用方程做预测和控制,并了解预测和控制的精度。多元回归分析预测法,是指通过对两个或两个以上的自变量与一个因变量的相关分析,建立预测模型进行预测的方法。当自变量与因变量之间存在线性关系时,称为多元线性回归分析。

MINITAB软件是现代质量管理统计的领导者,全球六西格玛实施的共同语言,它以无可比拟的强大功能和简易的可视化操作获得了广大质量学者和统计专家的青睐。MINITAB软件是为质量改善、教育和研究应用领域提供统计软件和服务,是质量管理和六西格玛实施软件工具,更是持续质量改进的良好工具软件。

1 多元线性回归分析的一般模型

多元线性回归分析的一般模型为:设 $x_1,x_2,\Lambda\ ,x_p$ 是 $p(\geq 2)$ 个自变量(解释变量),y 是因变量,多元线性回归模型的理论假设是

$$y=b_0+b_1x_1+b_2x_2+\Lambda\ +b_px_p+e\ ,$$

$$e\sim N(0,s^2)\ ,$$

其中,$b_0,b_1,b_2,\Lambda\ ,b_p$ 是 $p+1$ 个未知参数,b_0 称为回归常数,$b_1,b_2,\Lambda\ ,b_p$ 称为回归系数,$e\sim N(0,s^2)$ 为随机误差。

表1 父母身高与儿子身高

编号	父亲身高(cm) X_1	母亲身高(cm) X_2	年参加锻炼次数 X_3	儿子身高(cm) Y
1	172	163	90	176
2	171	159	70	172
3	169	158	50	170
4	171	161	65	174
5	167	159	50	169
6	172	163	100	177
7	172	160	60	171
8	170	162	70	173
9	175	166	110	182
10	179	166	100	183
11	176	164	90	180
12	171	159	80	174
13	167	158	60	172
14	176	163	70	177
15	172	162	70	175
16	181	169	90	186
17	174	167	80	182
18	170	161	70	174
19	183	169	120	187
20	176	165	110	182

表2 回归系数显著性检验表

项	系数	系数标准误	T 值	P 值
常量	-23.7	18.9	-1.25	0.228
父亲身高	0.303	0.137	2.22	0.042
母亲身高	0.880	0.181	4.85	0.000
锻炼次数	0.0593	0.0215	2.76	0.014

表3 ANOVA分析表

来源	自由度	Adj SS	Adj MS	F值	P 值
回归	3	527.139	175.713	140.14	0.000
父亲身高	1	6.159	6.159	4.91	0.042
母亲身高	1	29.521	29.521	23.54	0.000
锻炼次数	1	9.578	9.578	7.64	0.014
误差	16	20.061	1.254		
合计	19	547.200			

2 MINITAB软件建立模型

下面通过一个实例来详细讲解,如何运用MINITAB软件进行多元线性回归。现抽取20个家庭调查资料的部分变量,数据见表1,试对父母身高与儿子身高进行回归分析。

使用MINITAB软件,输入表1中数据,选择指令“统计>回归>回归”,在出现界面输入相应的变量名;打开“图形”窗,选择“四合一”及在“残差与变量”中填入各自变量名称;打开“存储”窗,选择“残差”、“标准化残差”及“拟合值”,点击“确定”后,得到输出结果。

MINITAB输出结果:

回归方程:

儿子身高=-23.7+0.303父亲身高+0.880母亲身高+0.0593锻炼次数

S=1.11974 R-sq=96.33% R-sq(调整)=95.65%

回归方程拟合出来以后,我们要解决以下几个问题:(1)给出方程显著性检验,从总体上判定回归方程有效与否。(2)给出方程总效果好坏的度量。(3)在回归方程效果显著时,对各个回归系数进行显著性检验,将效应不显著的自变量删除,以优化模型,这点在多元回归中尤为重要。(4)残差诊断,检验数据是否符合回归的基本假定,检验整个回归模型与数据拟合的是否很好,可否进一步改进回归方程来优化现有模型。

3 MINITAB输出结果分析

如何判断整个回归方程是否有意义?就要进行回归方程显著性检验,也就是要检验下列问题:H_0:模型无意义,H_1模型有意义。本例(表3)ANOVA表中 $P=0<0.05$,所以拒绝H_0:模型无意义,接受H_1模型有意义。说明在显著性水平a=0.05下,线性回归方程总效果是显著的。

如果实际观测值与拟合出来的回归线

(下转24页)

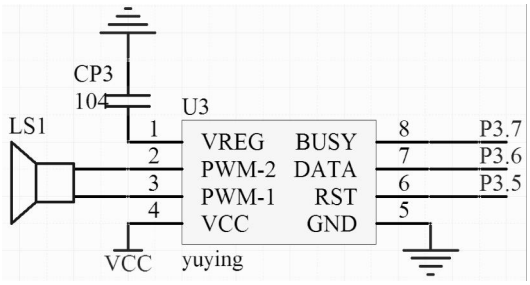

图4 报警电路

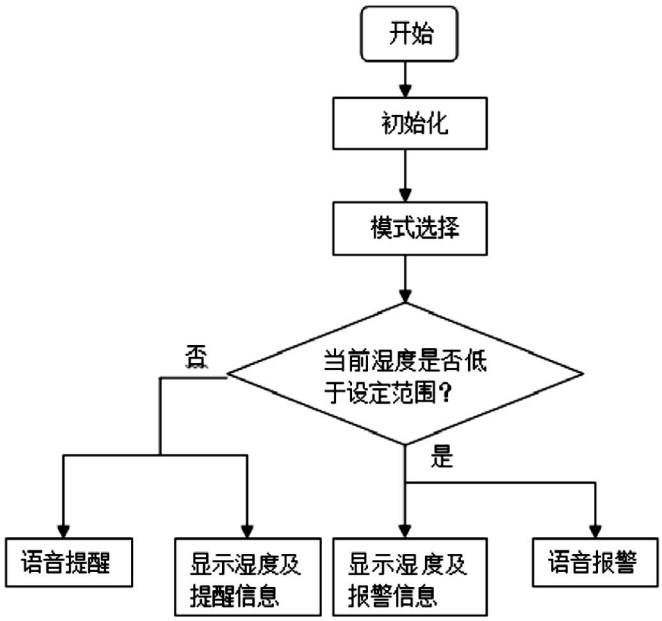

图5 系统程序流程图

范围也不同,根据盆栽所需土壤的合适湿度范围,可将盆栽大致分为湿生花卉、中生花卉、耐旱花卉三种。模式选择模块用于选择所监测盆栽的湿度类型,从而确定该盆栽的湿度监测范围。该部分电路用按键实现。

2.4 湿度显示模块

湿度显示模块用于显示当前湿度值,以及警示信息。该设计中的显示模块采用带中文字库的12864LCD液晶显示屏,如图3,它是一种具有4位/8位并行、2线或3线串行多种接口方式,内部含有国标一级、二级简体中文字库的点阵图形液晶显示模块;其显示分辨率为128×64,内置8192个16*16点汉字,和128个16*8点ASCII字符集。不论硬件电路结构或显示程序都要简洁得多,且该模块的价格也较低。

2.5 报警电路

报警电路采用语音芯片直接驱动喇叭的方式,用于实时播报当前湿度,以及土壤湿度低于设定湿度范围时的语音警报,由单片机控制其输出报警信号(图4)。

3 软件设计

该系统软件部分采用C语言编程,首先进行系统初始化,模式选择后确定湿度设定范围,检测当前湿度值与设定范围进行比较,如果在范围内,则输出湿度值及文字、语音提醒;若低于设定值,则输出湿度值并发出文字、语音报警信息,及时提醒为盆栽浇水,程序流程图如图5所示。

4 结语

该设计用单片机控制技术指导操作者科学地为盆栽浇水,使盆栽照料工作变得更加轻松愉快。系统采用集成了AD转换模块的单片机作为控制核心,并采用液晶显示模块显示提醒及报警信息,简化了硬件电路,降低了电路板的体积,而且操作方便。

参考文献

[1] 方泽鹏,黄双萍,陈仲涛.基于单片机的花盆土壤湿度控制系统设计[J].现代农业装备,2013(4):41-45.
[2] 张玮,王东锋.基于AT89S51单片机的微型土壤湿度检测仪设计[J].机电产品开发与创新,2010(7):74-75.
[3] 侯殿有.单片机C语言程序设计[M].北京:人民邮电出版社,2010.
[4] 郭天祥.新概念51单片机C语言教程入门、提高、开发、拓展全攻略[M].电子工业出版社,2009.

(上接22页)

很接近,就说明回归线与数据拟合的很好,就可以说回归方程的总效果很好。(表2)我们通常用R_{sq}、$R_{sq(adj)}$、S作为回归方程总效果的度量,以此来比较几种回归方程效果的好坏。R_{sq}是回归平方和占离差平方和的比率,其数值越接近1代表模型拟合的越好。当然R_{sq}并不是回归模型拟合效果的最好度量指标,因为当多一个自变量加入模型时,不管这个自变量是否显著,回归平方和就会增大,R_{sq}也会增大,这样就看不出新增加的自变量是否有意义,这点在多元回归中更为明显。因此我们用$R_{sq(adj)}$去修正R_{sq},以考虑总项数给模型带来的影响。$R_{sq(adj)} \leq R_{sq}$两者数值越接近越好,另一个指标是残差标准差$S = \sqrt{MS_E}$,它是从观察值与拟合回归线的平均偏离程度来度量的,也是回归模型中标准差σ的估计值。对于几个不同的回归方程的效果加以比较时,S是个最重要的指标,那个S最小,哪个回归方程就最小。

从本例输出结果看R_{sq}96.33%,$R_{sq(adj)}$=95.65%来看,两者很接近,S=1.11974比较小,模型还可以。

回归方程显著时,做回归系数显著性检验,一般假设H_0:β=0,H_1:β≠0,若$P<0.05$,则回归系数不为零,说明系数对应的自变量是显著的。当只有一个自变量时,回归方程显著性检验与回归系数检验是等价的,但是当自变量不止一个时,回归总效果显著不能排除某几个变量是无意义的。我们进行回归方程系数检验的目的,就是要找出是否有"滥竽充数"的自变量,把这些多余的自变量从方程中删除掉,以修正现有模型。

从本例输出结果看到三个自变量P值都小于0.05,故三个都为显著因子。

综上所述:我们认为模型为

$$y=-23.7+0.303x_1+0.880x_2+0.0593x_3$$

模型中,X_1系数0.303表示:如果父亲比同一代人的平均身高多1cm,那么他的儿子将比儿子那一代人的平均身高多出0.303 cm;X_2的系数解释也是如此;X_3的系数表示参加体育锻炼的次数和身高之间存在正相关;常数项一般没有与它相对应的实际意义上的解释。

参考文献

[1] 张海燕.基于多元线性回归模型的四川农村居民收入增长分析[J].统计观察,2010(13):88-90.
[2] 孙雪飞.回归分析在房地产销售中的应用[J].科技咨询导报,2007(26):168-169.
[3] 马逢时.六西格玛管理统计指南[M].北京:中国人民大学出版社,2012.