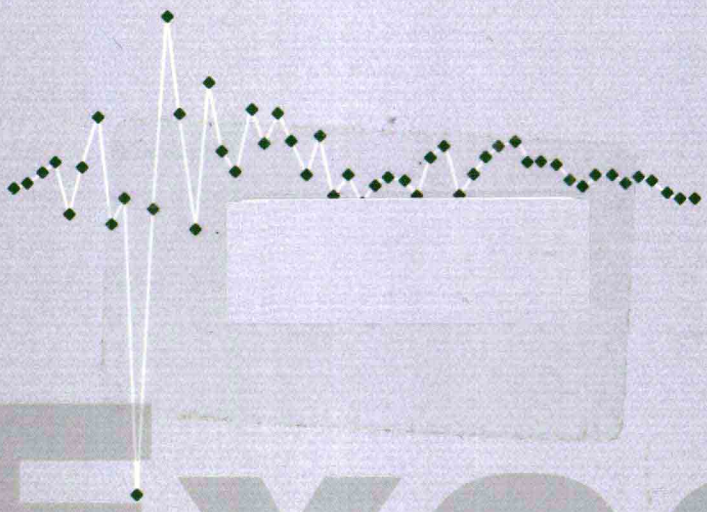
Mathematical Methods for Geography

基于Excel的地理数据分析

陈彦光 ◎ 编著



Excel

科学出版社
www.sciencep.com

基于 Excel 的地理数据分析

陈彦光　编著

国家科技部科技基础工作专项重点资助项目
地理学方法研究(2007FY140800)资助出版

科学出版社
北京

内 容 简 介

本书面向地理问题，基于 Excel 软件，叙述大量数学方法的应用思路和过程。内容涉及回归分析、主成分分析、聚类分析、判别分析、时(空)间序列分析、Markov 链、R/S 分析、线性规划、层次分析、灰色系统 GM(1，N)建模和预测方法等。通过模仿本书介绍的计算过程，读者可以加深对有关数学方法的认识和理解，并且掌握很多 Excel 的应用技巧。

这本书虽然是以地理数据为分析对象展开论述，但所涉及的内容绝大多数为通用方法。只要改变数据的来源，书中论述的计算流程完全可以应用到其他领域。

本书的初稿和修改稿先后在北京大学城市与环境专业研究生中试用八年，可供地理学、生态学、环境科学、地质学、经济学、城市规划学乃至医学、生物学等领域的学生、研究人员和工程技术人员阅读和参考。

图书在版编目(CIP)数据

基于 Excel 的地理数据分析 / 陈彦光编著. —北京：科学出版社，2010

ISBN 978-7-03-027182-2

Ⅰ.①基… Ⅱ.①陈… Ⅲ.①电子表格系统，Excel—应用—地理信息系统—数据—分析—教材 Ⅳ.①P208-39

中国版本图书馆 CIP 数据核字(2010)第 061128 号

责任编辑：韩 鹏 朱海燕 赵 冰 / 责任校对：赵桂芬

责任印制：钱玉芬 / 封面设计：王 浩

科学出版社出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

源海印刷有限责任公司印刷

科学出版社发行 各地新华书店经销

*

2010 年 4 月第 一 版 开本：787×1092 1/16

2010 年 4 月第一次印刷 印张：18

印数：1—3 000 字数：409 000

定价：48.00 元(含光盘)

(如有印装质量问题，我社负责调换)

前　言

要想成功地掌握一门数学方法，至少要熟悉如下几个环节：一是基本原理，即一种方法的理论基础和逻辑过程；二是应用范围，任何一种方法都有其自身的特长和功能局限，认识其优势和不足，才能真正有效地运用；三是算法或者运算规则系统，即一种为在有限步骤内解决数学问题而建立的可重复应用的计算流程体系；四是计算过程，即在一种方法的适用范围内，给定一组观测数据，并借助一定的算法获取所要求的计算结果；五是典型实例，即一种数学方法应用于现实问题的具体案例。如果还想进一步加深对一种数学方法的了解，还有第六个环节，那就是不同方法的融会贯通。

目前，我们学习绝大多数数学方法的基本原理都要求读者具备良好的高等数学知识，包括微积分、线性代数和概率与数理统计。不过，高等数学知识仅仅是掌握一门数学方法的必要条件。有了高等数学知识，我们就可以比较透彻地了解一种数学方法的逻辑结构，从而明确其内在原理。掌握一种方法的基本原理，大体上可以懂得其适用范围和功能局限。可是，所有这些，仅仅限于理论层面。要想借助相应的算法将一种数学原理有效地应用于现实问题，学会计算过程是非常关键的一个环节。任何一个数学方法的应用者，只有打通这个环节，才能在方法的运用方面尽可能地扬长避短。计算过程和典型实例是相辅相成的，典型实例是计算过程的结果，计算过程通常借助典型实例来显示其技术路线。

以最基本的数学方法——回归分析为例，学习该方法涉及如下过程。在基本思想方面，回归建模就是用数学语言刻画一组变量与某个变量之间的相关关系或者因果关系。关系的强弱通过回归系数表现，回归分析的核心问题就是模型参数值的估计。为此，需要一种有效的算法。目前的回归分析算法主要采用误差平方和最小的方法，即最小二乘法。在这个过程中，首先要采用线性方程组进行描述，理论上用到线性代数的知识；其次寻求误差平方和最小时的参数估计结果，理论上用到微积分的条件极值方法；在回归结果检验过程中，涉及误差的正态分布思想，这在理论上又用到大量的概率论和统计学原理。可是，虽然很多读者明白上述道理，但在具体应用过程中依然觉得似是而非。究其原因，主要在于不了解计算过程，没有掌握简明易懂的计算范例。

笔者编著本书的目的，就是帮助读者循序渐进地掌握一些数学方法的计算过程和简明范例，通过这个过程进一步加深对有关数学原理和方法的理解以及应用领域的认识，进而将不同的方法有机联系起来。全书的内容分为四大部分：一是相关分析和回归分析，主要讲述线性回归和逐步回归的计算过程；二是多元统计分析（以协方差逼近技术为主），主要讲述主成分分析、聚类分析和判别分析的计算过程；三是时空过程分析，包括时（空）间序列分析和时空随机过程分析，主要讲述自相关分析、自回归分析、周期图分析、功率（波）谱分析、Markov 链分析和 R/S 分析；四是系统分析，主要讲述层次分析（AHP）法、线性规划求解和灰色系统的建模与预测分析方法。

虽然书中讲到大量的有关 Excel 的应用技巧，但这不是一本关于 Excel 应用方法的

教材，而是基于 Excel 软件的数据处理和数学方法应用的教材。每一章的写作都采用相同的模式，即围绕一个或者若干个简明的例子，全方位地讲解一种数学方法的计算过程。书中讲述的有些数学方法处理过程是很实用的，如一元和多元回归分析方法、非线性回归建模方法、自回归分析方法、功率谱分析方法、Markov 链方法、AHP 法、线性规划求解方法、GM(1,1)和 GM(1, N)建模与预测方法等。也就是说，通过上述内容的学习，读者可以直接借助 Excel 处理实际工作中遇到的有关数据处理问题。另有一部分方法的讲述并不实用，而属于纯粹教学性质。逐步回归分析方法、主成分分析方法、聚类分析方法、判别分析方法、自相关分析方法等都属于此类。这些方法的计算过程烦琐，当数据量较大时，在 Excel 里开展工作速度缓慢而且容易出错。还有一些方法是介于上述两种情形之间的，包括周期图分析方法、R/S 分析方法等。当数据量较小时，可以采用这些方法在 Excel 里解决问题；但当数据量较大时，就得借助其他大型数学计算软件(如 Matlab、Mathcad)或者统计分析软件(如 SAS、SPSS)了。

读者可能产生疑问：既然一些方法在 Excel 里并不实用，为什么还要不厌其烦地讲述它们？这就回到前面提到的数学方法应用中的计算过程问题。笔者撰写本书的初衷不完全在于实用，大部分内容的实用性仅仅是本书内容的附带功能。笔者真正希望的，是借助本书实现如下目标：读者通过模仿一些计算过程，掌握有关模型建设的实例，进而理解有关数学方法的技术路线。以主成分分析方法为例，采用大型统计分析软件 SPSS，可以很方便地获得全面的计算结果。但是，SPSS 是一个“傻瓜”型软件，其计算过程对读者而言完全是一个“黑箱”。按照固定的程序操作该软件，不需要多少数学知识，就可以完成有关的统计计算。但是，如果不了解一种方法的计算过程，不知道这些方法的基本原理，即便 SPSS 输出结果，读者也没有办法给出准确的计算结果解释。如果读者首先在 Excel 里完成一个简明例子的计算，通过这个过程熟悉主成分分析的数学运算过程，然后再利用 SPSS 开展有关的数据整理和分析，就会主动和透明多了。当然，在阅读本书的过程中，读者会掌握 Excel 的很多功能和应用技巧，这些功能和技巧在未来的数据处理和分析过程中将会非常实用。

需要特别强调的是线性回归分析方法。这种方法非常简单而且基本，以致很多读者不重视该方法的深入学习和广泛练习。实际上，越是简单和基本的数学方法，使用频率越高，应用范围越广。一些复杂的数学方法，如主成分分析、判别分析、自回归分析、功率谱分析、小波分析、神经网络分析、灰色系统建模和预测分析等，都可以借助线性回归分析快速入门。本书讲述了基于回归分析的判别分析建模、自回归建模、周期图建模、R/S 分析建模、GM(1,1)和 GM(1, N)建模和预测等，并且在主成分分析等方法中应用了回归分析。这样，采用一种简明易懂的数学方法将多种数学方法贯通起来，读者可以通过回归分析了解多种数学方法的理论建设要点。

这部著作最初是北京大学研究生地理数学方法辅助教材，先后在北京大学原城市与环境学系、原环境学院、城市与环境学院试用八年。这不是简单的编写成果，而是带有很强的著作成分。实际上，在写作过程中，笔者参考的图书非常有限。最频繁使用的一部参考书是一本关于 Excel 函数的工具书——《Excel 2000 函数图书馆》，当然还有 Excel 自身附带的“帮助”内容。了解了 Excel 的数据分析、规划求解和数值拷贝功能之后，笔者所做的工作就是寻找合适的教学案例，根据相关的数学原理，在 Excel 中一步一步展开计算，

试读结束：需要全本请在线购买： www.ertongbook.com

并且详细地记录这些计算和分析过程。现在贡献给读者的，就是笔者对这些计算过程记录的整理结果。Excel 的常用函数功能、数值拷贝功能、数据分析和规划求解功能，加上笔者有关的数学方法原理方面的知识，以及相关案例的数据，就是这本书的主要写作源泉。

本书的写作特点是，借助简单的例子，从头到尾完整地演示各种数学方法的计算过程和分析思路。读者学习本书的方法则是，静下心来，从前到后重复一下笔者的计算过程，然后寻找一个类似的例子，自己按部就班模仿一遍。在模仿中学习，在思考中消化。通过阅读和操作，可以打开一些数学方法的“黑箱”，了解其内部结构，从而更好地进行运算结果的解读。然后，就可以借助 Excel 或者有关统计/数学软件处理自己研究的现实问题了。原则上，本书的每一章都相对独立，如果读者对 Excel 的基本功能比较熟悉，从任何一个部分都可以开始学习。但是，如果读者对 Excel 的基本功能不太熟悉，那就建议先系统学习第 1 章（一元线性回归分析）和第 2 章（多元线性回归分析）。然后再任选其他章节阅读。特别是本书第 1 章，笔者对 Excel 的有关功能和用法交代得非常详尽，对回归分析结果解释得相当细致。通过前面两章的学习和思考，读者基本上可以掌握 Excel 的常用数据分析操作技能。

最后对本书的一些数据处理和模型表现方式给出必要的说明。第一，数据处理过程前后是连贯的一体。对于任何一个案例，如果下一步用到上一步的数值计算结果，就直接调用有关结果所在的单元格，而不是重新输入近似值。这样做可以尽可能地降低数字出错的概率，并且提高计算的精度。但是，在行文表达的过程中，往往根据具体情况保留不同的小数位，绝大多数是根据 Excel 显示的结果给出计算值，有时根据具体情况有所变通。因此，书中显示的数字与实际计算用到的数字精度不相同。第二，模型的表达主要采用与 Excel 显示的公式接近的表现形式。数学工具的精确性主要体现在逻辑推理方面。当我们将一种数学方法应用于具体问题的时候，只能借助于某种算法估计模型参数。因此，数学模型的理论表达与经验表达是不一样的。由于如下两个方面的原因，我们对有关模型表达形式不作严格区分：一是尽可能与 Excel 公式显示结果保持一致；二是尽可能简化形式，贯通实质的数学过程。教学经验表明，过于严谨的数学形式反而不便于初学者学习数学方法。第三，统计检验标准采用默认的显著性水平。所有统计检验通过与否都是对于某个具体的置信度而言的。系统默认的显著性水平是 0.05，即置信度取 95%。采用这个临界值的好处在于，只要知道自由度，就可以方便地估计标准误差范围。因此，若无特别交代，书中所谓“检验通过”的显著性水平一律取 0.05。另外，对于运行前输入数据的 Excel 插图，尽可能将数据整理清晰，不压字；对于运行后输出结果的 Excel 插图，则采用系统默认的显示结果，虽然图中有些位置会因压字导致字符显示不全，但是便于读者操作过程中对照，也不会影响阅读理解。

光盘附带有各章使用的原始数据，读者可以调用这些数据，重复笔者给出的各种计算过程。此外，还有相关系数检验、F 检验、t 检验、Durbin-Watson 检验、卡方检验和调和分析的 Fisher 检验临界值表，供读者参考和使用。

作　者

2008 年 8 月

目　　录

第 1 章　一元线性回归分析

回归分析是最为基本的定量分析工具，很多表面看来与回归分析无关并且似乎难以理解的数学方法，可以借助回归分析得到简明的解释。通过回归分析，可以更好地理解因子分析、判别分析、自回归分析、功率谱分析、小波分析、神经网络分析等。在本书中，笔者将会建立回归分析与因子分析、判别分析、时间序列分析、灰色系统的 GM(1，N)预测分析等数学联系。在各种回归分析方法中，一元线性回归最为基本。熟练掌握这一套分析方法对学习其他数学工具非常有用。下面借助简单的实例详细解析基于 Excel 的一元线性回归分析。

【例】某地区最大积雪深度和灌溉面积的关系。为了估计山上积雪融化后对山下灌溉的影响，在山上建立观测站，测得连续 10 年的最大积雪深度和灌溉面积数据。利用这些观测数据建立线性回归模型，就可以借助提前得到的积雪深度数据，预测当年的灌溉面积大小。原始数据来源于苏宏宇等编著的《Mathcad 2000 数据处理应用与实例》。

1.1　模型的初步估计

这是非常初步的操作，却是非常重要的操作。我们在建立回归分析模型的过程中，首先要进行一些基本的试验。在 Excel 中，回归试验应用最为频繁的方法就是下面即将讲到的模型快速估计方法。

第一步，录入数据。数据录入结果如图 1-1-1。

第二步，作散点图。如图 1-1-2 所示，选中数据(包括自变量和因变量)，点击“图表向导”图标；或者在“插入”菜单中打开“图表(H)”。图表向导的图标为  。选中数据后，屏幕显示数据会变色(图 1-1-2)。

	A	B	C
1	年份	最大积雪深度（米）	灌溉面积（千亩）
2	1971	15.2	28.6
3	1972	10.4	19.3
4	1973	21.2	40.5
5	1974	18.6	35.6
6	1975	26.4	48.9
7	1976	23.4	45.0
8	1977	13.5	29.2
9	1978	16.7	34.1
10	1979	24.0	46.7
11	1980	19.1	37.4

图 1-1-1　数据录入结果

本书相关项目单位与此表中相同。1 亩≈667m^2

	A	B	C
1	年份	最大积雪深度（米）	灌溉面积（千亩）
2	1971	15.2	28.6
3	1972	10.4	19.3
4	1973	21.2	40.5
5	1974	18.6	35.6
6	1975	26.4	48.9
7	1976	23.4	45.0
8	1977	13.5	29.2
9	1978	16.7	34.1
10	1979	24.0	46.7
11	1980	19.1	37.4

图 1-1-2　选中作图的数据序列

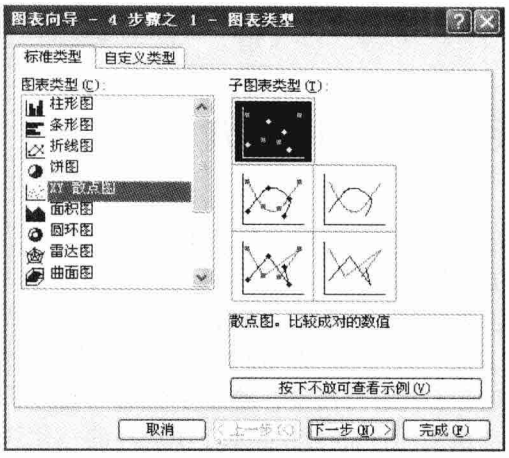


图 1-1-3　图表向导中的散点图选项

点击“图表向导”以后，弹出如下对话框(图 1-1-3)。在左边的“图表类型(C)”栏中选中“XY 散点图”，点击“完成”按钮，立即出现散点图的原始形式(图 1-1-4)。

第三步，模型估计。这一过程又可以细分为如下几个步骤。

(1) 选中散点：用鼠标指向图 1-1-4 中的数据点列，点击右键，出现如图 1-1-5 的选择菜单。

(2) 添加趋势线：点击“添加趋势线(R)”，弹出如图 1-1-6 的选择框。

(3) 选项设置：在分析“类型”中选择“线性(L)”，然后打开选项单(图 1-1-7)。

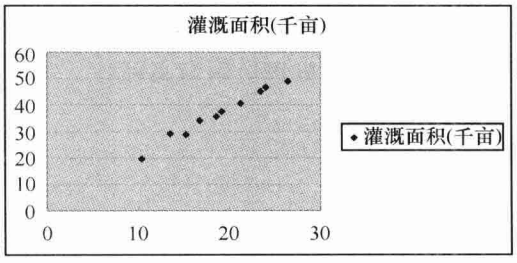


图 1-1-4　Excel 给出的散点图

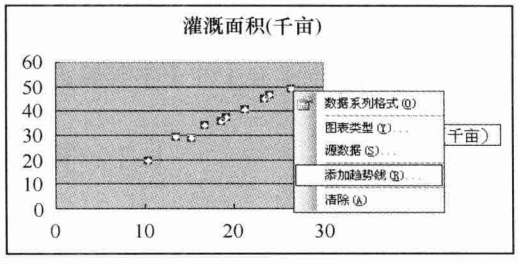


图 1-1-5　选中图中的散点系列

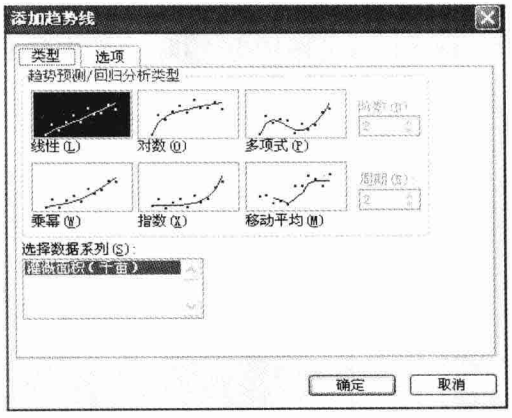


图 1-1-6　添加线性趋势线的选项

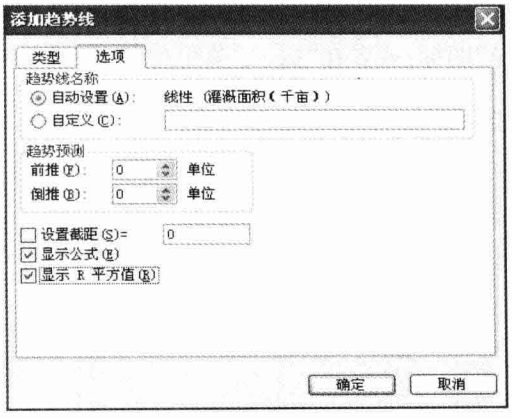


图 1-1-7　添加趋势线的选项设置

试读结束：需要全本请在线购买：　www.ertongbook.com

(4) 获取结果：在选择框中选中“显示公式(E)”和“显示 R 平方值(R)”(图 1-1-7)，确定，立即得到回归结果如图 1-1-8。

在图 1-1-8 中，给出了回归模型和相应的测定系数即拟合优度。模型为 $y=2.3564+1.8129x$，相关系数平方为 $R^2=0.9789$。

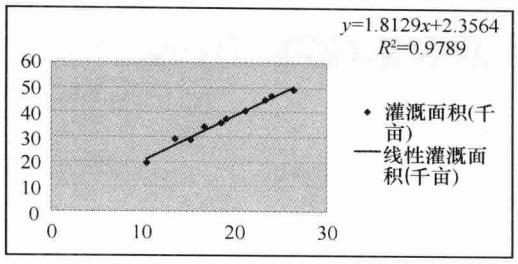


图 1-1-8　模型的初步估计结果

1.2　详细的回归过程

1.2.1　回归建模

回归模型的快速估计过程非常简便，但结果也过于简略。除了模型的截距、斜率估计结果和相关系数平方等统计量之外，没有其他方面的统计信息。为了对模型参数估计值开展深入的统计分析，我们需要掌握详细的回归分析过程。在 Excel 中，回归分析过程可以分为若干步骤完成，第一、二步与 1.1 节“模型的初步估计”给出的步骤完全一样，姑且从略。下面从第三步讲起。

观察如图 1-1-4 所示的散点图，判断点列分布是否具有线性趋势。只有当数据具有线性分布特征时，才能采用线性回归分析方法。从图中可以看出，本例数据形成的散点呈现线性分布特征，可以进行线性回归。详细步骤如下。

图 1-2-1　选中“数据分析”

1. 打开对话框

点击“工具”下拉菜单，可见数据分析选项(图 1-2-1)。

双击“数据分析(D)”选项，弹出“数据分析”选项框(图 1-2-2)。

2. 回归分析选项

在图 1-2-2 中选择“回归”，确定，弹出如图 1-2-3 的选项表。

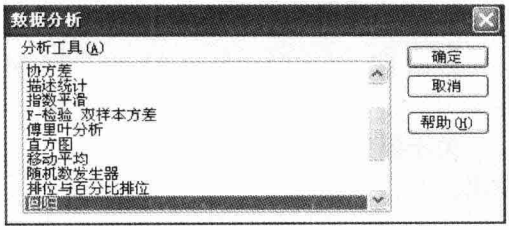


图 1-2-2　数据分析选项框

第一种选择方式：包括数据标志。X、Y 值的输入区域(B1:B11,C1:C11)，标志，置信度(95%)，新工作表组，残差，线性拟合图(图 1-2-4)。

第二种选择方式：不包括数据标志。X、Y 值的输入区域(B2:B11,C2:C11)，置信度(95%)，新工作表组，残差，线性拟合图(图 1-2-5)。

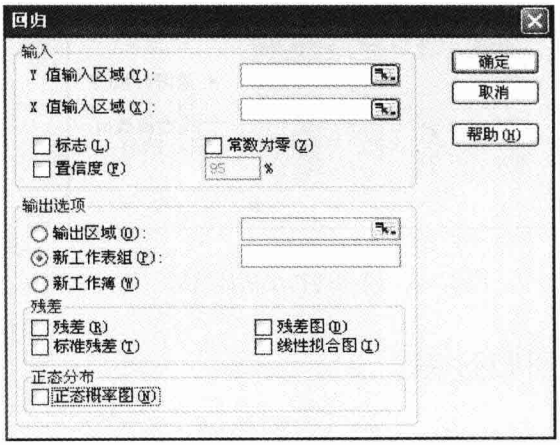


图 1-2-3　数据分析选项框

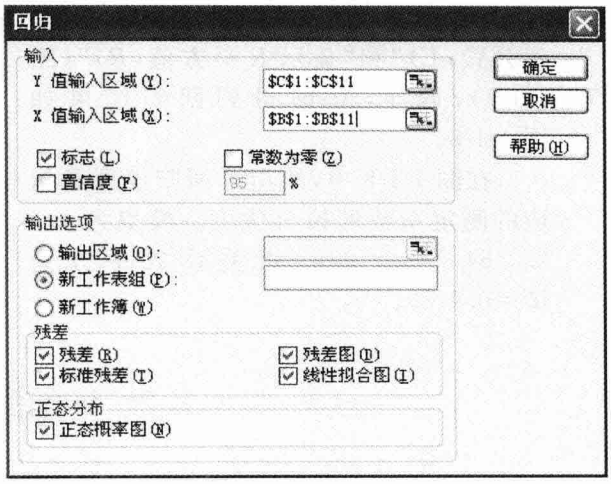


图 1-2-4　包括数据标志

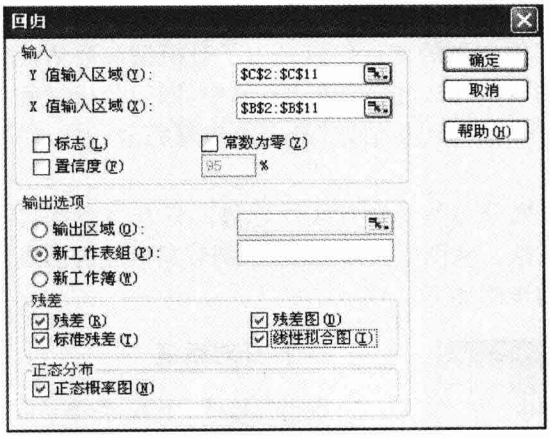


图 1-2-5　不包括数据标志

注意：选中数据“标志”和不选“标志”，X、Y 值的输入区域是不一样的。前者包括数据标志“最大积雪深度(米)”和“灌溉面积(千亩)”，后者不包括。当在输入栏的数据范围中包括数据标志所在的单元格时，必须选择“标志”选项，否则不能选中“标志”。这一点务必注意。

3. 给出回归结果

设置完成以后，确定，即可得到全部回归结果(图 1-2-6)。

4. 读取参数

在图 1-2-6 所示的结果中，读取如下数据，据此建立模型并开展统计分析。截距：$a=0.356$；斜率：$b=1.813$；相关系数：$R=0.989$；测定系数：$R^2=0.979$；F 值：$F=371.945$；t 值：$t=19.286$；回归标准误差：$s=1.419$；回归平方和：SSr=748.854；剩余平方和：SSe=16.107；总平方和：SSt=764.961。

5. 写出模型表达式

根据上面的回归结果建立回归模型，并对结果进行检验。回归模型为

$$\hat{y}=a+bx=2.356+1.813x$$

关于模型的统计检验，R、R^2、F 值、t 值、标准误差值等均可以直接从回归结果中读出。

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	回归统计								
4	Multiple	0.989416							
5	R Square	0.978944							
6	Adjusted	0.976312							
7	标准误差	1.418924							
8	观测值	10							
9									
10	方差分析								
11		df	SS	MS	F	gnificance F			
12	回归分析	1	748.8542	748.8542	371.9453	5.42E-08			
13	残差	8	16.10676	2.013345					
14	总计	9	764.961						
15									
16		Coefficien	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
17	Intercept	2.356438	1.827876	1.289167	0.233363	-1.85865	6.571527	-1.85865	6.571527
18	最大积雪	1.812921	0.094002	19.28588	5.42E-08	1.596151	2.029691	1.596151	2.029691
19									
20									
21									
22	RESIDUAL OUTPUT					PROBABILITY OUTPUT			
23									
24	观测值	灌溉面积（	残差	标准残差		百分比排位	面积（千亩）		
25	1	29.91284	-1.31284	-0.98136		5	19.3		
26	2	21.21082	-1.91082	-1.42836		15	28.6		
27	3	40.79036	-0.29036	-0.21705		25	29.2		
28	4	36.07677	-0.47677	-0.35639		35	34.1		
29	5	50.21755	-1.31755	-0.98489		45	35.6		
30	6	44.77879	0.221209	0.165356		55	37.4		
31	7	26.83087	2.369128	1.770947		65	40.5		
32	8	32.63222	1.46778	1.097181		75	45		
33	9	45.86654	0.833457	0.623017		85	46.7		
34	10	36.98323	0.41677	0.31154		95	48.9		

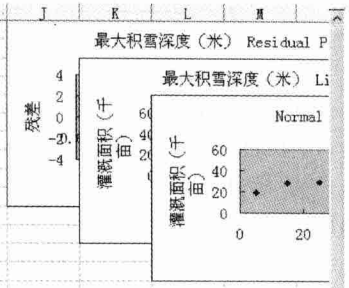


图 1-2-6　线性回归结果

1.2.2　模型的统计检验

对于一元线性回归，只需要开展相关系数检验、标准误差检验和 DW 检验。不过，作为方法介绍，不妨给出较为全面的说明。

1. 相关系数检验

在相关系数检验表中查出，当显著性水平取 $\alpha=0.05$、剩余自由度为 $df=10-1-1=8$ 时，相关系数的临界值为 $R_{0.05,8}=0.632$。显然

$$R=0.989\ 416>0.632=R_{0.05,8}$$

检验通过。有了 R 值，F 值和 t 值均可计算出来。

2. F 检验

F 值的计算公式和结果为

$$F=\frac{R^2}{\frac{1}{n-m-1}(1-R^2)}=\frac{0.989\ 416^2}{\frac{1}{10-1-1}(1-0.989\ 416^2)}=371.945>5.318=F_{0.05,1,8}$$

式中：$n=10$ 为样品数目；$m=1$ 为自变量数目。显然与表中的结果一样。

3. t 检验

t 值的计算公式和结果为

$$t=\frac{R}{\sqrt{\frac{1-R^2}{n-m-1}}}=\frac{0.989\,416}{\sqrt{\frac{1-0.989\,416^2}{10-1-1}}}=19.286>2.306=t_{0.05,8}$$

可见，F 值为 t 值的平方，即有 $19.286^2=371.945$。上述结果 Excel 都已经直接给出，这里通过验算有助于理解这些统计量之间的联系。

查 F 分布表和 t 分布表，可以得到 F 值和 t 统计量的临界值。实际上，在 Excel 中，利用公式 finv 可以方便地查出 F 统计量的临界值。语法是：finv(α，m，$n-m-1$)，即

finv(显著性水平，自变量数目，样品数目－自变量数目－1)

任选一个单元格，输入公式“=FINV(0.05,1,10－2)”，回车，立即得到 5.317 645，即

$$F_{0.05,1,8}=5.318$$

类似地，利用公式 tinv 可以方便地查出 t 统计量的临界值。语法是：tinv(α，$n-m-1$)，即

tinv(显著性水平，样品数目－自变量数目－1)

任选一个单元格，输入公式“=TINV(0.05,10－2)”，回车，立即得到 2.306 005 6，即

$$t_{0.05,8}=2.306$$

4. 标准误差检验

回归结果中给出了残差(图 1-2-7)，据此可以计算标准误差。首先求残差的平方

$$\varepsilon_i^2=(y_i-y_i)^2$$

然后求残差平方和

$$\mathrm{SSe}=\sum_{i=1}^{n=10}\varepsilon_i^2=1.724+\cdots+0.174=16.107$$

于是标准误差为

$$s=\sqrt{\frac{1}{n-m-1}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}=\sqrt{\frac{1}{\nu}\mathrm{SSe}}=\sqrt{\frac{16.107}{8}}=1.419$$

从而得到变异系数

$$\frac{s}{\bar{y}}=\frac{1.419}{36.53}=0.0388<10\%\sim15\%=0.1\sim0.15$$

利用平方和函数 sumsq 可以直接求出残差平方和。如图 1-2-7 所示，残差序列位于第三列，即 C 列(图 1-2-6)。在任意空白单元格输入公式“=SUMSQ(C24:C34)”，回车，得到 16.106 760 4。用这个数除以剩余自由度 8，然后开平方根，即可得到标准误差 1.418 924。

5. DW 检验

DW 值的计算公式及结果为

$$\mathrm{DW}=\frac{\sum_{i=2}^{n}(\varepsilon_i-\varepsilon_{i-1})^2}{\sum_{i=1}^{n}\varepsilon_i^2}=\frac{(-1.911+1.313)^2+\cdots+(0.417-0.833)^2}{(-1.313)^2+(-1.911)^2+\cdots+0.417^2}=0.751$$

在 Excel 中计算 DW 值非常方便。只要在图 1-2-4 所示的选项中选中“残差(R)”，最后的回归结果就会给出残差序列。将残差序列复制出来，在适当的地方粘贴两次，注意最

好错位粘贴(图 1-2-8)。然后,利用其中一列减去另外一列,得到残差之差,或者残差序列的差分。注意,原来的数据有 $n=10$ 个,残差之差应为 $n-1=9$ 个。利用函数 sumsq 计算残差序列的平方和,结果为 16.1068;然后用 sumsq 计算残差之差的平方和,结果为 12.0949。第二个平方和除以第一个平方和,即残差之差的平方和除以残差的平方和,即可得到 DW 值。具体说来,DW 值=12.0949/16.1068=0.751。

24	观测值	预测 灌溉	残差	标准残差	残差平方
25	1	29.91284	-1.31284	-0.98136	1.723544
26	2	21.21082	-1.91082	-1.4283553	3.651222
27	3	40.79036	-0.29036	-0.2170504	0.084312
28	4	36.07677	-0.47677	-0.3563903	0.227309
29	5	50.21755	-1.31755	-0.9848852	1.735949
30	6	44.77879	0.221209	0.16535611	0.048933
31	7	26.83087	2.369128	1.77094725	5.612766
32	8	32.63222	1.46778	1.09718082	2.154379
33	9	45.86654	0.833457	0.62301729	0.69465
34	10	36.98323	0.41677	0.31153965	0.173697
35				残差平方和	**16.1068**

图 1-2-7　y 的预测值及其相应的残差等

		残差		
观测值	残差	-1.31284	残差之差	
1	-1.31284	-1.91082	-0.59798	
2	-1.91082	-0.29036	1.620453	
3	-0.29036	-0.47677	-0.18641	
4	-0.47677	-1.31755	-0.84078	
5	-1.31755	0.221209	1.538763	
6	0.221209	2.369128	2.147919	
7	2.369128	1.46778	-0.90135	
8	1.46778	0.833457	-0.63432	
9	0.833457	0.41677	-0.41669	
10	0.41677			DW值
	16.1068		**12.0949**	**0.75092**

图 1-2-8　利用残差计算 DW 值

最后是 DW 统计检验。取 $\alpha=0.05$,$n=10$,$m=1$(剩余自由度 df=10−1−1=8),在统计表中查表得 $d_l=0.94$,$d_u=1.29$。显然,DW 值=0.751 $<d_l=0.94$,这意味着有序列正相关,预测的结果可能令人怀疑。当然,对于本例,由于 $n<15$,DW 值的估计结果不可靠,严格意义的 DW 检验无法进行。

1.3　回归结果详解

1.3.1　数据表的解读

利用 Excel 的数据分析进行回归,可以得到一系列的统计参量。下面将图 1-2-6 回归结果摘要(summary output)放大(图 1-3-1)。

下面逐步说明如下。

1. 回归统计表

这一部分给出了相关系数、测定系数、校正测定系数、标准误差和样品数目(图 1-3-2)。

逐行解释如下:

(1) Multiple 对应的数据是相关系数(correlation coefficient),即 $R=0.989\ 416$。

(2) R Square 对应的数值为测定系数(determination coefficient),或称拟合优度(goodness of fit),它是相关系数的平方,即有 $R^2=0.989\ 416^2=0.978\ 944$。

(3) Adjusted 对应的是校正测定系数(adjusted determination coefficient),校正公式为

$$R_a=1-\frac{(n-1)(1-R^2)}{n-m-1}=1-\frac{(n-1)(1-R^2)}{\nu}$$

SUMMARY OUTPUT

回归统计	
Multiple	0.989416
R Square	0.978944
Adjusted	0.976312
标准误差	1.418924
观测值	10

方差分析					
	df	SS	MS	F	gnificance F
回归分析	1	748.8542	748.8542	371.9453	5.42E-08
残差	8	16.10676	2.013345		
总计	9	764.961			

	Coefficien	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	2.356438	1.827876	1.289167	0.233363	-1.85865	6.571527	-1.85865	6.571527
最大积雪	1.812921	0.094002	19.28588	5.42E-08	1.596151	2.029691	1.596151	2.029691

图 1-3-1　回归结果摘要(局部放大)

回归统计	
Multiple	0.989416
R Square	0.978944
Adjusted	0.976312
标准误差	1.418924
观测值	10

图 1-3-2　回归统计表

式中:n 为样品数;m 为变量数;ν 为自由度(df);R^2 为测定系数。对于本例,$n=10$,$m=1$,$R^2=0.978\ 944$,代入上式得

$$R_a = 1-\frac{(10-1)(1-0.978\ 944)}{10-1-1} = 0.976\ 312$$

(4) 标准误差(standard error)对应的即回归标准误差,计算公式上一节已经给出。将 SSe=16.106 76 代入计算公式可得

$$s = \sqrt{\frac{1}{10-1-1}\times 16.106\ 76} = 1.418\ 924$$

这个结果在前面进行过验算。

方差分析					
	df	SS	MS	F	gnificance F
回归分析	1	748.8542	748.8542	371.9453	5.42E-08
残差	8	16.10676	2.013345		
总计	9	764.961			

图 1-3-3　方差分析表(ANOVA)

(5) 观测值对应的是样品数目(这里为年数),即有 $n=10$。

2. 方差分析表

方差分析部分包括自由度、误差平方和、均方差、F 值、t 统计量、P 值、参数估计结果的变化范围等(图 1-3-3)。

逐列、分行解释如下:

第一列 df 对应的是自由度(degree of freedom)。

(1) 第一行是回归自由度 dfr,等于自变量数目,即 dfr=m。

(2) 第二行为剩余自由度 dfe,或者残差自由度,等于样品数目减去自变量数目再减去 1,即有 dfe=$n-m-1$。在计算公式中,剩余自由度通常用 ν 表示。

(3) 第三行为总自由度 dft,等于样品数目减 1,即有 dft=$n-1$。对于本例,$m=1$,$n=10$,因此,dfr=1,dfe=$n-m-1=8$,dft=$n-1=9$。

显然,三者的关系是

回归自由度+剩余自由度=总自由度

第二列 SS 对应的是误差平方和,或称变差。

(1) 第一行为回归平方和或称回归变差 SSr,即有

试读结束:需要全本请在线购买: www.ertongbook.com