第3章

电子商务数据分析方法

【章节目标】

1．了解静态指标和动态指标的含义
2．掌握相关分析的计算过程
3．重点掌握一元线性回归的计算过程，了解多元线性规划和非线性回归的计算过程
4．重点掌握移动平均和指数平滑两种时间序列预测模型
5．了解聚类分析等数据挖掘算法的主要计算过程

【学习重点、难点】

1．正确区分电子商务主要数据指标的分类
2．准确计算相关系数
3．计算线性回归模型的参数并建立回归模型
4．对时间序列数据进行移动平均和指数平滑处理
5．K-means 聚类算法的计算步骤与应用

【案例导入】

神奇的购物篮分析

在一家超市中，人们发现了一个特别有趣的现象：尿布与啤酒这两种风马牛不相及的商品居然摆在一起。但这一奇怪的举措居然使尿布和啤酒的销量大幅增加了。这可不是一个笑话，而是一直被商家津津乐道的发生在美国沃尔玛连锁超市的真实案例。原来，美国的妇女通常在家照顾孩子，所以她们经常会嘱咐丈夫在下班回家的路上为孩子买尿布，而丈夫在买尿布的同时会顺手购买自己爱喝的啤酒。这个发现为商家带来了大量的利润，但是如何从浩如烟海又杂乱无章的数据中发现啤酒销售和尿布销售之间的联系呢？这又给了我们什么样的启示呢？这个案例说明，通过分析商品大数据，利用能够找出商品之间关联关系的数据分析算法，进一步提取客户的购买行为，能够为商业决策提供辅助的决策支持。

案例来源：高勇．啤酒与尿布：神奇的购物篮分析[M]．北京：清华大学出版社，2008.

3.1　统计分析

在电子商务数据分析与应用的实践中，针对不同的分析目的，分析方法不尽相同，但其主要包括统计分析与数据挖掘两大类。其中，统计分析方法包括静态分析指标、动态分析指标、统计指数、抽样推断、相关分析与回归分析等内容。

3.1.1　静态分析指标

静态分析指标是用来说明社会经济现象数量特征的。由于社会经济现象及其发展的复杂性，静态分析指标呈现多样性，可以将其归纳为 4 类：总量指标、相对指标、平均指标和变异指标。

1. 总量指标

总量指标是反映社会经济现象在一定时间、地点和条件下的总体规模或水平的统计指标。它的表现形式为绝对数，故又称为统计绝对数。例如，某家淘宝店铺的总营业额、员工总数、产品销售总量等，都是反映社会经济现象总量的，均可视为总量指标。如表 3.1 所示，2020 年 10 月的店铺总交易量 399 单可视为总量指标。

表 3.1　2020 年 10 月某网店 11 名员工的工资收入与当月交易量

员 工 编 号	年龄/岁	工资/元	交易量/单
1001	28	4800	30
1002	33	5200	45
1003	28	4800	29
1004	26	5000	43
1005	25	4900	30
1006	23	4800	28
1007	28	5000	32
1008	23	4900	31
1009	32	5200	44
1010	28	5000	42
1011	34	5500	45
合计		55 100	399

2. 相对指标

两个有联系的统计指标的比率称为相对指标。与总量指标伴随有量纲单位不同，相对指标在绝大多数情况下采用无名数标识。无名数是一种抽象化的数值，多用倍数、系数、成数、百分数等表示。例如，2018 年某天猫店铺的总营业额为 2017 年的 125.7%，如图 3.1 所示，该店铺与同行业其他店铺相比，其物流服务评分高于同行业平均分 53.82%。

3. 平均指标

平均指标是同类社会经济现象总体各单位的某一数量标志在一定时间、地点和条件下的数量差异抽象化的代表性水平指标，其数据表现为平均数。平均指标可以反映社会经济现象总体的综合特征，也可以反映各变量值分布的集中趋势。平均指标按计算和确定的方法不同，可分为算术平均数、调和平均数、众数和中位数等。例如，某天猫店铺员工的平均工资、某店铺平均评价值等。如图 3.1 所示，该天猫店铺的描述相符平均分为 4.9、服务态度平均分为 4.9、物流服务平均分为 4.9。

图 3.1　某天猫店铺与同行业其他店铺的对比数据

【温馨提示】

算术平均数、调和平均数、众数和中位数的应用

2020 年 10 月某网店 11 名员工的工资收入与当月交易量如表 3.1 所示。

（1）算术平均数。例如，计算本月全部员工的月平均工资，则 11 名员工的月工资总额为 55 100 元，月平均工资为 5009 元（55 100/11）。

（2）调和平均数。例如，计算编号为 1001～1005 号员工的本月平均交易量，公式为 5/(1/30+1/45+1/29+1/43+1/30)=34.1，即这 5 名员工的本月平均交易量为 34.1 单。

（3）众数。例如，计算该网店员工年龄的一般水平。经过汇总可知，在该网店 11 名员工中，25、26、32、33、34 岁各 1 名，23 岁 2 名，28 岁 4 名，因此，28 是该网店员工年龄的众数。

（4）中位数。例如，计算该网店员工年龄的中位数。首先将年龄数据从小到大排列，由于该组数据由 11 个数据组成，因此选择排在第 6 位的员工年龄 28 作为中位数。

4. 变异指标

变异指标是综合反映总体各单位标志值变异程度的指标。它显示了总体中变量数值分布的离散趋势，是说明总体特征的另一种重要指标，与平均数的作用相辅相成。变异指标按计算的方法不同分为极差、四分位差、平均差、标准差和方差等。

【温馨提示】

极差、四分位差、平均差、标准差和方差的应用

仍采用如表 3.1 所示的表格数据分析该网店员工的工资收入变异指标。

（1）极差。工资极差=最大工资额-最小工资额=5500-4800= 700 元。

（2）四分位差。四分位差是上四分位数（Q3，位于 75%）与下四分位数（Q1，位于 25%）

的差。首先将 11 名员工的工资收入从小到大排列，Q1 的位置是 3，对应 4800 元；Q3 的位置是 9，对应 5200 元，因此四分位差为 5200−4800=400 元，表明该网店有 50%的员工的月工资收入在 4800 元～5200 元，最大差异为 400 元。

（3）平均差。月工资收入的平均差为∑(每位员工的工资收入−月平均工资)/员工总数=1745.5/11=158.64 元。

（4）标准差和方差。方差是各数据与其算术平均数的离差二次方的平均值，而方差的平方根即标准差。因此，本例题中，月工资收入的方差为 42 644.64，标准差为 206.51 元。

3.1.2　动态分析指标

动态分析又称时间数列分析，主要用来描述和探索现象随时间发展变化的数量规律，对处于不断发展变化的社会经济现象，从动态的角度进行分析。通过对以下内容的学习，可利用各种动态分析指标对社会经济现象进行分析。

1. 动态数列

动态数列指将同类指标在不同时间上的数值按时间的先后顺序排列起来而形成的统计数列，又称为时间数列。它是一种常见的经济数据表现形式。

2. 动态数列分类

动态数列主要分为以下 3 类。

（1）绝对数动态数列：把一系列同类的总量指标按时间先后顺序排列起来而形成的动态数列。例如，某网店 2020 年 10 月 1 日—2020 年 10 月 5 日的访客数量形成绝对数动态数列，如表 3.2 所示。

（2）相对数动态数列：把一系列同类的相对指标按时间先后顺序排列起来而形成的动态数列。例如，某网店 2020 年 10 月 1 日—2020 年 10 月 5 日的支付转化率形成相对数动态数列，如表 3.2 所示。

（3）平均数动态数列：把一系列同类的平均指标按时间先后顺序排列起来而形成的动态数列。例如，某网店 2020 年 10 月 1 日—2020 年 10 月 5 日的平均客单价形成平均数动态数列，如表 3.2 所示。

表 3.2　某网店 2020 年 10 月 1 日—2020 年 10 月 5 日的动态数列

指　　标	2020 年 10 月 1 日	2020 年 10 月 2 日	2020 年 10 月 3 日	2020 年 10 月 4 日	2020 年 10 月 5 日
访客数/人	980	1201	1120	789	902
支付转化率/%	5.54	6.89	4.99	6.02	8.14
平均客单价/元	33.98	45.75	56.25	29.76	48.25

【知识拓展】

典型的电子商务数据指标分类

电子商务数据很复杂，数据来源渠道也多样化，因此电子商务数据分析的指标也很多。在

本书的附录 A 中，流量指标属于静态指标，转化指标属于动态指标。请同学们参照附录 A 中对各指标的具体解释，考虑其他电子商务数据指标中还有哪些属于静态指标、哪些属于动态指标。

3.1.3 统计指数

统计指数分析法是经济分析中广泛应用的一种方法。其中，最具代表性的就是关于物价指标的编制，即用现行价格与过去价格的对比来反映价格的变化情况，后来过渡到综合反映多种商品价格的变动情况。

1. 统计指数的作用

统计指数在社会经济领域应用广泛，这是因为统计指数具有独特的功能，能够发挥重要的作用，具体表现在以下几个方面。

（1）综合反映了复杂社会经济现象总体在时间和空间方面的变动方向和变动程度。这是统计指数最重要的作用。在社会经济现象中存在着大量不能直接加总或不能直接对比的复杂总体，为了反映和研究它们的变动方向和变动程度，只有编制统计指数才能得到解决。

（2）分析和测定了社会经济现象总体变动受各因素变动的影响。在社会经济现象总体中包含着数量因素和质量因素，通过编制数量因素指数和质量因素指数，可以分析和测定各因素变动对总体变动的影响。

（3）研究了平均指标变动及其受水平因素和结构因素变动的影响。平均指标中包含水平因素和结构因素，因此可以编制可变组成指数、不变组成指数和结构影响指数，从而研究平均指标的变动及其各因素变动对平均指标变动的影响。

2. 统计指数的类型

按不同的研究目的和要求，统计指数可进行以下分类。

1）个体指数和总指数

按研究对象的范围不同，统计指数可分为个体指数和总指数。个体指数反映某种社会经济现象中个别事物变动的情况，如某一种商品物价的变动情况；总指数则综合反映某种事物包括若干个别事物总的变动情况，如若干商品总的物价变动情况。有时为了研究需要，在介于个体指数与总指数之间，还编制了组指数（或类指数），组指数的编制方法与总指数相同。

2）数量指标指数和质量指标指数

按表示的特征不同，统计指数可分为数量指标指数和质量指标指数。数量指标指数反映社会经济现象总体的规模和水平的变动状况，如产量指数、职工人数指数等；质量指标指数反映社会经济现象总体内涵质量的变动，如商品物价指数、劳动生产率指数等。

3）动态指数和静态指数

统计指数按其本来的含义，都是指动态指数，但在实际运用过程中，其含义渐渐推广到了静态事物和空间对比，因此产生了静态指数。静态指数指在同一时间条件下，对不同单位、不同地区的同一事物数量进行对比形成的指数；或者对同一单位、同一地区的计划指标与实际指标进行对比形成的指数。

4）定基指数和环比指数

按在指数数列中采用的基期不同，统计指数可分为定基指数和环比指数。定基指数指在数列中以某一固定时期的水平作为对比基准的指数；环比指数指以其前一时期的水平作为对比基准的指数。

5）综合指数和平均指数

按研究方法不同，统计指数可分为综合指数和平均指数。综合指数是将不可直接度量的指数化指标通过同度量因素转化为可以合计的总量指标，然后将不同时期的总量指标进行对比，以综合反映社会经济现象的动态变化。平均指数是以个体指数为基础，通过简单平均或加权平均的方法计算总指数。这两种指数是独立的指数形式，且存在内在的联系。

【知识拓展】

电子商务发展指数

2019 年 5 月，2019 中国国际大数据产业博览会期间，《中国电子商务发展指数报告（2018）》正式发布[①]。报告认为，近年来我国电子商务发展取得了规模影响持续扩大、体系逐步完善、法制环境不断健全等成就。同时，报告从规模、成长、渗透、支撑 4 个方面对各省电子商务的发展水平进行了综合测评。

（1）规模指数：反映电子商务发展的市场规模，主要考查各省电子商务交易额、网络零售额、有电子商务活动的企业数等指标。该指数值越高，表明该地区电子商务的市场规模越大。规模指数可以反映当前电子商务市场自身的发展水平。

（2）成长指数：反映电子商务发展的成长水平，主要通过增长率考查各省在电子商务交易与零售额等方面的表现。该指数值越高，表明该地区电子商务成长潜力越大。成长指数可以反映电子商务发展的未来预期。

（3）渗透指数：反映各省电子商务对经济活动的影响程度，表明电子商务对传统经济发展的影响。该指数值越高，表明该地区的经济活动中电子商务渗透程度越高，电子商务对传统产业的影响也越大。

（4）支撑指数：反映各省与电子商务发展相关的保障能力，主要考查各省在电子商务相关的物流设施、人力资本及技术环境等方面的建设情况。该指数值越高，表明该地区电子商务的发展环境越好。

基于对上述 4 个指数的综合分析，广东省连续四年领跑全国，先导省份包括广东、浙江、北京、上海和江苏。省级电商梯队基本形成，后发省份突围存在困难。电子商务规模整体呈现“东强西弱”的局势，集群效应显现。

▶▶ 3.1.4　抽样推断

抽样推断（Sample Inference）是在抽样调查的基础上，利用样本的实际资料计算样本指标，并据此推算总体相应数量特征的一种统计分析方法。统计分析的主要任务是反映社会经济现象总体的数量特征。但在实际工作中，不可能、也没有必要每次都对总体的所有单位进行全面调

① 资料来源：中华人民共和国国家互联网信息办公室，http://www.cac.gov.cn/2019-05/29/c_1124554997.htm。

查。在很多情况下，只需抽取总体的部分单位作为样本，然后通过分析样本的实际资料来估计和推断总体的数量特征，以达到对社会经济现象总体的认识。

1. 抽样推断的作用

抽样推断的作用主要包括：在无法进行全面调查或进行全面调查有困难时，可采用抽样调查来推断总体特征；采用抽样调查可以节省费用和时间，以提高调查的时效性和经济效果；可用来对全面资料进行检验和修正；可用于工业生产过程的质量控制；可通过对某种总体的假设进行检验来判断这种假设是否正确，以决定行动的取舍。

2. 抽样推断的内容

1）全及总体和样本总体

全及总体是研究对象，样本总体是观察对象，两者是既有区别又有联系的不同范畴。全及总体又称为母体，简称总体，指所要认识的、具有某种共同性质的许多单位的集合体。样本总体又称为子样，简称样本，指从全及总体中随机抽取出来的、代表全及总体的那部分单位的集合体。样本的单位数称为样本容量，通常用小写英文字母 n 来表示。随着样本容量的增大，样本对总体的代表性越来越高，并且当样本单位数足够多时，样本平均数接近总体平均数。例如，针对 100 万名淘宝用户，随机抽取 1000 名进行网络购物满意度调查，则 100 万名淘宝用户就是总体，而被抽中的 1000 名淘宝用户则构成样本。

2）总体参数和样本统计量

总体参数又称为全及指标，是根据总体各单位的标志值或标志属性进行计算，并反映总体某种属性或特征的综合指标。常用的总体参数有总体平均数（或总体成数）、总体标准差（或总体方差）。样本统计量又称为样本指标，是由样本各单位标志值计算出来反映样本特征，并用来估计总体参数的综合指标（抽样指标）。样本统计量是样本变量的函数，用来估计总体参数，因此与总体参数相对应。样本统计量有样本平均数（或抽样成数）、样本标准差（或样本方差）。

3）样本容量和样本个数

通常将样本容量不少于 30 个的样本称为大样本，不及 30 个的样本称为小样本。社会经济统计的抽样调查多属于大样本调查。样本个数又称为样本可能数目，指从一个总体中可能抽取的样本个数。一个总体有多少样本，则样本统计量就有多少种取值，从而形成样本统计量的分布，此分布是抽样推断的基础。

4）重复抽样和不重复抽样

重复抽样指从总体单位中抽取一个单位进行观察，记录后再放回总体中，然后抽取下一个单位，连续抽取样本。不重复抽样指从总体单位中抽取一个单位进行观察，记录后不再放回总体中，然后在余下的总体中抽取下一个单位。

3. 抽样推断在电子商务数据分析中的应用

假设某网店有 4 名物流人员，每人的日出库量分别为 40、50、70、80 件。先随机抽取 2 人，分别采用重复抽样和不重复抽样的方式，计算样本统计量。

首先根据重复抽样和不重复抽样形成样本，如表 3.3 所示。在重复抽样的条件下，样本平均数的平均数为 960/16=60 件，样本平均误差为$(2000/16)^{1/2}$=11.18 件。在不重复抽样的条件下，

样本平均数的平均数为 720/12=60 件，样本平均误差为$(1000/12)^{1/2}$=9.13 件。

表 3.3　重复抽样和不重复抽样的样本内容

样本单位：件

序　号	重 复 抽 样			不重复抽样		
	样本变量 x_1	样本平均数	离差平方和	样本变量 x_2	样本平均数	离差平方和
1	40，40	40	400	/	/	/
2	40，50	45	225	40，50	45	225
3	40，70	55	25	40，70	55	25
4	40，80	60	0	40，80	60	0
5	50，40	45	225	50，40	45	225
6	50，50	50	100	/	/	/
7	50，70	60	0	50，70	60	0
8	50，80	65	25	50，80	65	25
9	70，40	55	25	70，40	55	25
10	70，50	60	0	70，50	60	0
11	70，70	70	100	/	/	/
12	70，80	75	225	70，80	75	225
13	80，40	60	0	80，40	60	0
14	80，50	65	25	80，50	65	25
15	80，70	75	225	80，70	75	225
16	80，80	80	400	/	/	/
	合计	960	2000	合计	720	1000

3.2　相关分析与回归分析

相关分析与回归分析是数理统计中两种重要的统计分析方法，应用非常广泛。这两种方法从本质上来讲有许多共同点，它们都是从数据内在逻辑方面分析变量之间的联系的。相关分析是回归分析的基础和前提，只有当两个或两个以上的变量之间存在高度的相关关系时，进行回归分析寻求其相关的具体形式才有意义。

▶▶ 3.2.1　相关分析

1. 相关关系的概念

相关关系指变量之间存在的一种不确定的数量依存关系，即当一个变量的数值发生变化时，另一个变量的数值也相应地发生变化，但变化的数值不是确定的，而是在一定范围内的。例如，广告是提高销售量的重要手段，但广告投入不是销售量增加的唯一影响因素，产品的质量、价格、销售方式等都会对销售量产生影响。在研究广告投入与销售量的关系时，发现广告投入的增加一般会带来销售量的增长，但广告投入每增加一个固定的量，销售量并不是以确定的量增

加的，而是表现为一个随机变量。广告投入与销售量的这种关系就是相关关系。

在现实社会经济生活中，现象之间的这种相关关系是非常普遍的，一个变量的变化往往不止受到一个变量的影响。当我们在考查一个变量与其中一个影响变量的关系时，由于其他因素的存在，二者之间的量化关系不是完全确定的，而是带有随机的成分。统计研究的就是这种受随机因素影响而不能唯一确定的变量关系。

2. 相关关系的种类

1）按程度分类

（1）完全相关：两个变量之间的关系是一个变量的数量变化由另一个变量的数量变化唯一确定，即函数关系。

（2）不完全相关：两个变量之间的关系介于不相关和完全相关之间。

（3）不相关：如果两个变量彼此的数量变化互相独立，那么两个变量之间没有关系。

2）按方向分类

（1）正相关：两个变量的变化趋势相同，从散点图可以看出各点散布的位置是从左下角到右上角的区域，即当一个变量的值由小变大时，另一个变量的值也由小变大。

（2）负相关：两个变量的变化趋势相反，从散点图可以看出各点散布的位置是从左上角到右下角的区域，即当一个变量的值由小变大时，另一个变量的值由大变小。

3）按形式分类

（1）线性相关（直线相关）：当相关关系的一个变量变动时，另一个变量也相应地发生均等的变动。

（2）非线性相关（曲线相关）：当相关关系的一个变量变动时，另一个变量也相应地发生不均等的变动。

4）按变量数目分类

（1）单相关：只反映一个自变量和一个因变量的相关关系。

（2）复相关：反映两个及两个以上的自变量同一个因变量的相关关系。

（3）偏相关：当研究因变量与两个或多个自变量相关时，如果把其余的自变量看成不变的（当作常量），只研究因变量与其中一个自变量之间的相关关系，那么称为偏相关。

变量 X 和变量 Y 的相关关系示意图如图 3.2 所示。

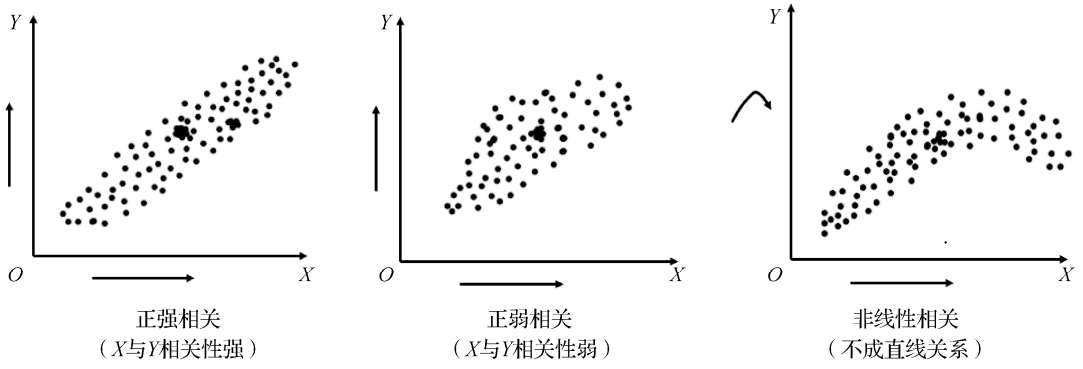
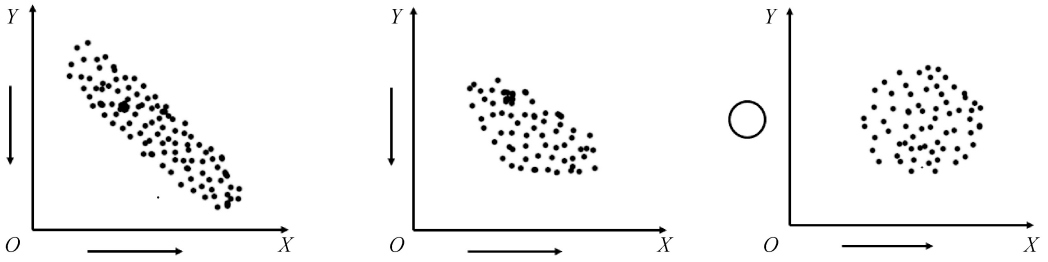


图 3.2　变量 X 和变量 Y 的相关关系示意图



负强相关
（X增大，Y迅速下降）

负弱相关
（X增大，Y缓慢下降）

不相关
（Y不随X增减）

图 3.2　变量 X 和变量 Y 的相关关系示意图（续）

3. 相关系数

相关系数 R 是描述变量 x 与 y 之间线性关系密切程度的一个数量指标。

$$R=\frac{\sum_{i=1}^{n}x_iy_i-n\overline{xy}}{\sqrt{\sum_{i=1}^{n}x_i^2-n\overline{x}^2}\sqrt{\sum_{i=1}^{n}y_i^2-n\overline{y}^2}}=\frac{l_{xy}}{\sqrt{l_{xx}l_{yy}}}\ (-1\leqslant R\leqslant 1) \tag{3-1}$$

式中，$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$，$\overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$，$l_{xx}=\sum_{i=1}^{n}(x_i-\overline{x})^2$，$l_{yy}=\sum_{i=1}^{n}(y_i-\overline{y})^2$，$l_{xy}=\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$。

当 R=1 时，表示完全正相关；当 R=−1 时，表示完全负相关；当 R=0 时，表示不相关。查相关系数临界值表，若 $R>R_\alpha(n-2)$，则线性相关关系显著，通过检验，可以进行预测；反之，没有通过检验。若不查表，通过经验判断，则 R 的范围在 0.3～0.5 表示低度相关；R 的范围在 0.5～0.8 表示显著相关；R 的范围在 0.8 以上表示高度相关。

3.2.2　回归分析

1. 一元线性回归分析

一元线性回归分析是处理两个变量 x（自变量）和 y（因变量）之间关系的最简单模型，研究的是这两个变量之间的线性相关关系。

$$y_i=a+bx_i+u_i\ \ (i=1,2,\cdots,n) \tag{3-2}$$

式（3-2）称为一元线性回归模型（One Variable Linear Regression Model）。其中，u_i 是一个随机变量，称为随机项；a，b 两个常数可通过最小二乘法求得，称为回归系数（参数）；i 表示变量的第 i 个观察值，共有 n 组样本观察值。

2. 多元线性回归分析

多元线性回归模型（Multivariate Linear Regression Model）的基本假设是在对一元线性回归模型的基本假设基础之上，还要求所有自变量彼此线性无关，这样随机抽取 n 组样本观察值就

可以进行参数估计了。

$$y_i = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + u_i \quad (i = 1, 2, \cdots, n) \tag{3-3}$$

3. 非线性回归分析

在许多实际问题中，不少经济变量之间的关系为非线性的，可以通过变量代换把本来应该用非线性回归方式处理的问题近似转化为线性回归问题，再进行分析预测，如表 3.4 所示。

表 3.4　常见的非线性模型及线性变换的方式

非线性模型	函 数 形 式	变 换 方 式	线 性 模 型
幂函数形式	$y = a\,x^b$	$y' = \log_2 y$ $x' = \log_2 x$ $a' = \log_2 a$	$y' = a' + bx'$
双曲线形式	$1/y = a + b(1/x)$	$y' = 1/y$ $x' = 1/x$	$y' = a + bx'$
对数函数形式	$y = a + b\log_2 x$	$x' = \log_2 x$	$y = a + bx'$
指数函数形式	$y = a\mathrm{e}^{bx}$	$y' = \ln y$ $a' = \ln a$	$y' = a' + bx$
多项式曲线形式	$y = b_0 + b_1 x + b_2 x^2 + \cdots + b_k x^k$	$x_1 = x,\ x_2 = x^2 \cdots x_k = x^k$	$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$

▶▶ 3.2.3　相关分析与回归分析的应用

1. 案例数据

某网店通过付费流量进行推广，该网店的运营总监认为，网店的付费流量投入与用户访问量、网店利润是正相关的。同时，流量、访问量与网店利润的变化均存在一定联系。利用 Excel 对如表 3.5 所示的数据进行相关分析与回归分析。

表 3.5　某网店的运营数据

日　　期	付费流量投入/元	访问量/次	网店利润/元
2020/09/01	1659	3420	520.5
2020/09/02	1989	4662	522.9
2020/09/03	2195	4925	527.1
2020/09/04	2255	4831	531.5
2020/09/05	2329	5302	534.7
2020/09/06	2375	5535	537.4
2020/09/07	2364	5815	540.4
2020/09/08	2354	6348	543.2
2020/09/09	2418	6561	545.3
2020/09/10	2534	6644	551.5
2020/09/11	2568	6883	554.6
2020/09/12	2835	6844	557.9

2. 相关分析的操作

在 Excel 的“数据分析”模块中找到相关系数，单击“确定”按钮，如图 3.3 所示。如果未发现“数据分析”选项，那么通过单击“文件→选项→加载项→分析工具库”，再单击“确定”按钮，加载“数据分析”模块。

在打开的“相关系数”对话框中，单击“输入区域”右侧的折叠按钮，在工作表中选择数据区域“$B$1：$C$13”，设置分组方式为“逐列”，再单击输出区域“$B$15”，最后单击“确定”按钮，如图 3.4 所示。

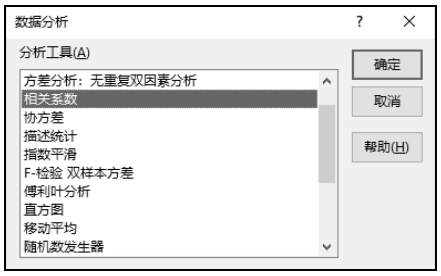


图 3.3　选择“相关系数”功能

图 3.4　设置“相关系数”的相关参数

单击“确定”按钮后，可在表格的“$B$15”区域得到结果，如图 3.5 所示，表明付费流量投入与网店利润之间存在正相关，相关系数约为 0.924，属于高度相关关系。

3. 回归分析的操作

1）利用 Excel 图表进行回归分析

单击“插入”选项卡，选择图表选项中的“XY（散点图）”，再单击散点图中“带平滑线的散点图”选项，如图 3.6 所示。

	付费流量投入/元	网店利润/元
付费流量	1	
网店利润	0.923723182	1

图 3.5　相关系数的结果

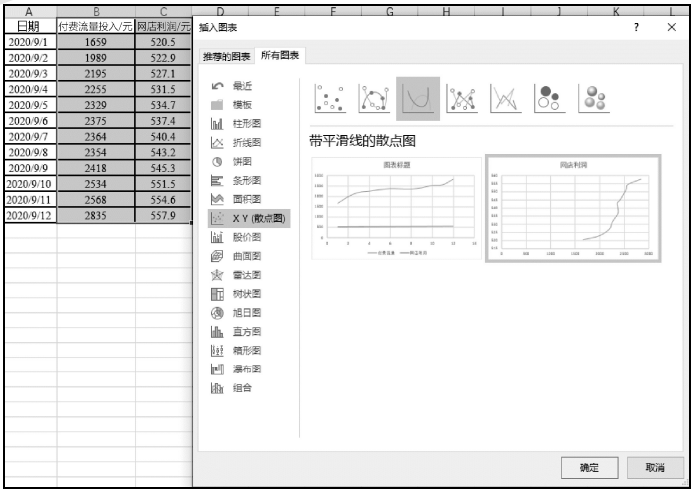


图 3.6　绘制散点图

对如图 3.6 所示的散点图进行优化处理，将横轴最小值设置为“1500”，如图 3.7 所示。

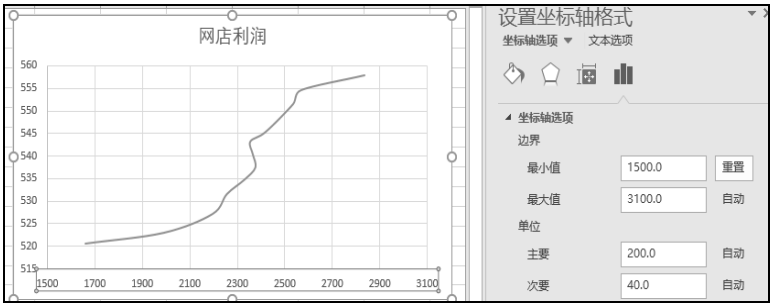


图 3.7　设置散点图的横坐标

右击如图 3.7 所示的曲线，选择“添加趋势线”，再单击“线性”单选按钮，勾选“显示公式”和“显示 R 平方值”复选框，如图 3.8 所示。

添加趋势线后，出现如图 3.9 所示的线性趋势线，则一元线性回归公式为 y=0.0382 x+450.07，R^2=0.8533。

图 3.8　添加趋势线的选项

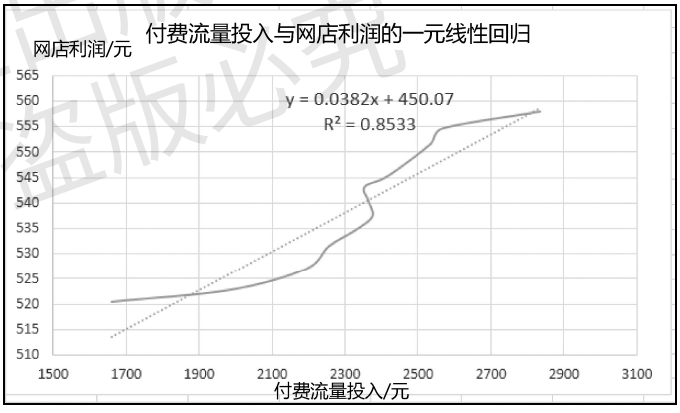


图 3.9　添加趋势线的效果图①

2）应用数据分析的回归功能

在数据分析的“分析工具”下拉列表框中选择“回归”选项，单击“确定”按钮，如图 3.10 所示。

网店利润为 Y 值，因此 Y 值输入区域设置为“$C$1：$C$13”；付费流量投入为 X 值，因此 X 值输入区域设置为“$B$1：$B$13”，置信度设置为 95%，计算结果的输出区域从“$B$15”开始。单击“确定”按钮，如图 3.11 所示。

① 图中的字母 x、y、R 应为斜体，但该公式为计算机自动生成，无法修改。

单击“确定”按钮后，出现如图 3.12 所示的一元线性回归分析结果，包括各参数值及模型检验的结果。图 3.12 中的阴影数据与趋势线结果相同。

	A	B	C
1	日期	付费流量投入/元	网店利润/元
2	2020/9/1	1659	520.5
3	2020/9/2	1989	522.9
4	2020/9/3	2195	527.1
5	2020/9/4	2255	531.5
6	2020/9/5	2329	534.7
7	2020/9/6	2375	537.4
8	2020/9/7	2364	540.4
9	2020/9/8	2354	543.2
10	2020/9/9	2418	545.3
11	2020/9/10	2534	551.5
12	2020/9/11	2568	554.6
13	2020/9/12	2835	557.9

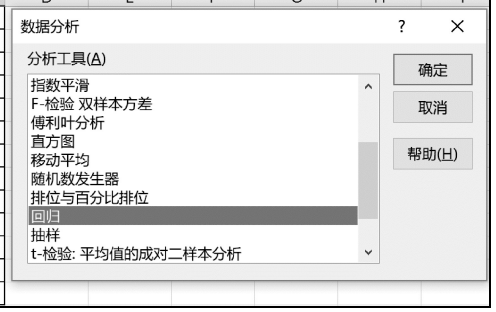


图 3.10　回归分析功能

图 3.11　一元线性回归分析选项

回归统计	
Multiple R	0.923723182
R Square	0.853264517
Adjusted R Si	0.838590969
标准误差	4.900178106
观测值	12

方差分析

	df	SS	MS	F	Significance F
回归分析	1	1396.279212	1396.279212	58.14984227	1.78726E-05
残差	10	240.1174547	24.01174547		
总计	11	1636.396667			

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	450.0745593	11.7360615	38.34971035	3.46729E-12	423.9249847	476.2241339	423.9249847	476.2241339
付费流量投入/元	0.03824593	0.005015462	7.625604387	1.78726E-05	0.027070784	0.049421076	0.027070784	0.049421076

图 3.12　一元线性回归分析结果

若要继续进行付费流量投入、访问量与网店利润之间的二元线性回归分析，则其操作过程与上述步骤基本相同，仅需要将 *X* 值输入区域设置为“$B$1:$C$13”、*Y* 值输入区域设置为“$D$1:$D$13”，如图 3.13 所示。

	A	B	C	D
1	日期	付费流量投入/元	访问量/次	网店利润/元
2	2020/9/1	1659	3420	520.5
3	2020/9/2	1989	4662	522.9
4	2020/9/3	2195	4925	527.1
5	2020/9/4	2255	4831	531.5
6	2020/9/5	2329	5302	534.7
7	2020/9/6	2375	5535	537.4
8	2020/9/7	2364	5815	540.4
9	2020/9/8	2354	6348	543.2
10	2020/9/9	2418	6561	545.3
11	2020/9/10	2534	6644	551.5
12	2020/9/11	2568	6883	554.6
13	2020/9/12	2835	6844	557.9

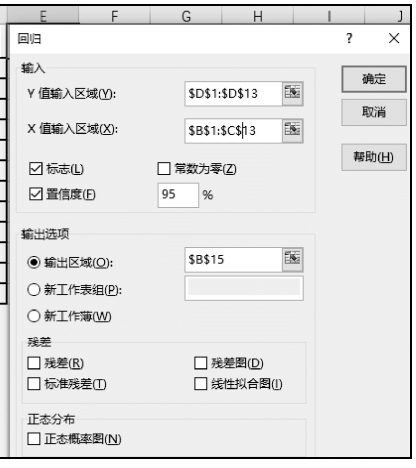


图 3.13　二元线性回归模型的参数设置

单击如图 3.13 所示的“确定”按钮，得到如图 3.14 所示的二元线性回归分析结果，则二元

线性回归模型的公式为 y=0.0123 x_1 +0.0078 x_2 +466.47，R^2=0.9232，检验值 F 为 54.14。其中，付费流量投入为 x_1，店铺访问量为 x_2，网店利润为 y。

回归统计								
Multiple R	0.9608677							
R Square	0.9232668							
Adjusted R Square	0.9062149							
标准误差	3.7352078							
观测值	12							
方差分析								
	df	SS	MS	F	ignificance F			
回归分析	2	1510.8307	755.41533	54.144737	9.603E-06			
残差	9	125.566	13.951778					
总计	11	1636.3967						
	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	466.47395	10.620032	43.923968	8.209E-12	442.44976	490.49813	442.44976	490.49813
付费流量投入/元	0.0123214	0.009822	1.2544745	0.2412688	-0.009897	0.0345404	-0.009897	0.0345404
访问量 /次	0.0077594	0.002708	2.8654018	0.0186157	0.0016336	0.0138852	0.0016336	0.0138852

图 3.14　二元线性回归分析结果

【数据视野】

大数据相关性大于因果性

传统的数据分析更多是追求因素关系的。例如，在进行实验设计时，通过控制某个变量的变化来评估其对相关结果的影响，以确定是不是这个变量产生的影响。在大数据时代，面对激烈的竞争环境，不过多地追求因果关系，更多是通过相关分析，找到事物之间的联系，再依据数据相关分析结果，快速做出决策。这也许会是企业取得竞争优势的基础。

虽然电子商务大数据可以直接通过顾客评论、退货、投诉等数据，很快地分析出顾客的流失特征，但是这些分析很难明确地回答一个顾客真正流失的原因。通过大数据可以分析得出有过差评的顾客流失可能性更高的结论，但这是流失顾客表现出来的一个特征，而真正导致顾客流失的原因是不是与物流也存在关系呢？很难下一个明确的结论，只能是猜测，但并不重要了，重要的是已经通过这些相关指标，识别出了谁可能会流失，从而产生事前预警，这才是商业分析的最大价值。重视相关性，不是不要因果关系，因果关系还是基础，科学的基石还是要的，只是在高速信息化的时代，为了得到即时信息，实时预测，在快速的大数据分析技术下，找到相关性信息，就可预测用户的行为，支持企业的快速决策。

3.3　时间序列分析

3.3.1　时间序列数据

随着计算机技术和大容量存储技术的发展及多种数据获取技术的广泛应用，人们在日常事务处理和科学研究中积累了大量数据。被保存的数据绝大部分都是时间序列类型的数据。所谓时间序列数据，就是按照时间先后顺序排列各观测记录的数据集。时间序列数据在社会生活的各个领域都广泛存在，如金融证券市场中每天的股票价格变化；商业零售行业中某项商品每天

的销售额；气象预报研究中某一地区每天的气温与气压的读数；在生物医学中某一症状的病人在每个时刻的心跳变化等。不仅如此，时间序列数据也是反映事物运动、发展、变化的一种最常见的图形化描述方式。

3.3.2　移动平均方法

1. 一次移动平均法

一次移动平均法是在算术平均法的基础上加以改进的，其基本思想是每次取一定数量周期的数据平均，再按时间顺序逐次推进。每推进一个周期，舍去前一个周期的数据，增加一个新周期的数据，再进行平均。一次移动平均法一般只应用于一个时期后的预测（预测第 t+1 期）。

一次移动平均数 $M_t^{(1)}=\dfrac{y_t+y_{t-1}+\cdots y_{t-N+1}}{N}$，$M_t^{(1)}$ 代表第 t 期一次移动平均值，N 代表在计算移动平均值时选定的数据个数。一般情况下，N 越大，修匀的程度越强，波动也越小；N 越小，对变化趋势反应越灵敏，但修匀的程度越差。在实际预测中，可以利用试算法，即选择几个 N 值进行计算，比较它们的预测误差，从中选择预测误差较小的 N 值。

2. 二次移动平均法

当时间序列具有线性增长的发展趋势时，用一次移动平均法预测会出现滞后偏差，表现为对线性增长的时间序列的预测值偏低。这时，可通过二次移动平均法来计算。二次移动平均法是将一次移动平均再进行一次移动平均，然后建立线性趋势模型。

二次移动平均法的线性趋势预测模型：

$$\hat{y}_{t+\tau}=\hat{a}_t+\hat{b}_t\tau \tag{3-4}$$

式中，截距为 $\hat{a}_t=2M_t^{(1)}-M_t^{(2)}$，斜率为 $\hat{b}_t=\dfrac{2}{N-1}\left(M_t^{(1)}-M_t^{(2)}\right)$，$\tau$为预测超前期。$M_t^{(1)}$ 为第 t 期一次移动平均值；$M_t^{(2)}$ 为第 t 期二次移动平均值，计算公式为 $M_t^{(2)}=\dfrac{M_t^{(1)}+M_{t-1}^{(1)}+\cdots+M_{t-N+1}^{(1)}}{N}$，$N$ 代表在计算移动平均值时选定的数据个数。

3.3.3　指数平滑方法

1. 一次指数平滑法

设时间序列为 $y_1, y_2, ..., y_t$，则一次指数平滑公式为

$$S_t^{(1)}=\alpha y_t+(1-\alpha)S_{t-1}^{(1)} \tag{3-5}$$

式中，$S_t^{(1)}$ 为第 t 期的一次指数平滑值；α为加权系数，$0<\alpha<1$。

2. 二次指数平滑法

当时间序列没有明显的变动趋势时，使用第 t 期一次指数平滑法就能直接预测第 t+1 期的值。但当时间序列的变动呈现直线趋势时，用一次指数平滑法来预测存在着明显的滞后偏差。

修正的方法是在一次指数平滑的基础上再进行一次指数平滑，利用滞后偏差的规律找出曲线的发展方向和发展趋势，然后建立直线趋势预测模型，即二次指数平滑法。

设一次指数平滑为 $S_t^{(1)}$，则二次指数平滑 $S_t^{(2)}$ 的计算公式为

$$S_t^{(2)} = \alpha S_t^{(1)} + (1-\alpha) S_{t-1}^{(2)} \tag{3-6}$$

若时间序列 $y_1, y_2, \ldots, y_t$ 从某时期开始具有直线趋势，且认为在未来时期也按此直线趋势变化，则其与趋势移动平均类似，可用以下直线趋势模型来预测：

$$\hat{y}_{t+T} = a_t + b_t T \quad (T = 1, 2, \cdots) \tag{3-7}$$

式中，t 为当前时期数；T 为由当前时期数 t 到预测期的时期数；$\hat{y}_{t+T}$ 为第 $t+T$ 期的预测值；a_t 为截距，b_t 为斜率，其计算公式为 $a_t = 2S_t^{(1)} - S_t^{(2)}$，$b_t = \dfrac{\alpha}{1-\alpha}(S_t^{(1)} - S_t^{(2)})$。

3.3.4 季节指数方法

1. 季节指数水平法

季节指数水平法指变量在一年内以季（月）的循环为周期特征，通过计算变量的季节指数来达到预测目的的一种方法。季节指数水平法的预测过程：首先分析判断时间序列数据是否呈季节性波动。通常可将 3～5 年的资料按月或季展开，绘制历史曲线图，通过观察其在一年内有无周期性波动来做出判断；然后将各种因素结合起来考虑，即考虑它是否还受长期趋势变动的影响、是否受随机波动的影响等。

季节指数水平法的计算步骤如下。

第一步：收集 3 年以上各年中各月或季数据 Y_t，形成时间序列。

第二步：计算各年同季或同月的平均值 $\overline{Y}_i$：$\overline{Y}_i = \sum_{i=1}^{n} Y_i / n$，$Y_i$ 为各年各月或各季观察值，n 为年数。

第三步：计算所有年度所有季或月的平均值 $\overline{Y}_0$：$\overline{Y}_0 = \sum_{i=1}^{n} \overline{Y}_i / n$，$n$ 为一年季数或月数。

第四步：计算各季或各月的季节比率 f_i（季节指数）：$f_i = \overline{Y}_i / \overline{Y}_0$

第五步：计算预测期趋势值 $\hat{X}_t$。趋势值是不考虑季节变动影响的市场预测趋势估计值，它的计算方法有多种，如可以以观察年的年均值除以一年的月数或季数。

第六步：建立季节指数水平预测模型，即 $\hat{Y}_t = \hat{X}_t \cdot f_t$。

2. 季节指数趋势法

季节指数趋势法指在时间序列观察值既有季节周期变化，又有长期趋势变化的情况下，首先建立趋势预测模型，然后在此基础上求得季节指数，最后建立数学模型进行预测的一种方法。

季节指数趋势法的计算步骤如下。

第一步：以一年的季数 4 或一年的月数 12 为 N，对观察值的时间序列进行 N 项移动平均。由于 N 为偶数，因此应对相邻两期的移动平均值再平均后对正，形成新序列 M_t，并以此为长期趋势。

第二步：将各期观察值除以同期移动平均值，得到季节比率 f_t（$f_t = Y_t / M_t$），以消除趋势。

第三步：将各年同季或同月的季节比率平均，季节平均比率 F_i 可消除不规则变动。i 表示某季度或月份。

第四步：计算时间序列线性趋势预测值 $\hat{X}_t$，模型为 $\hat{X}_t = a + bt$。这里可采用移动平均法：

$$b = \frac{M_t\text{末项} - M_t\text{首项}}{M_t\text{项数}}；\quad a = \frac{\sum_{t=1}^{n} Y_t - b\sum_{t=1}^{n} t}{n}。$$

第五步：求季节指数趋势预测值 $\hat{Y}_t = \hat{X}_t \cdot F_i$。

3.3.5　时间序列分析算法实例

已知某天猫店铺 2016～2019 年季度零售额数据，请对如表 3.6 所示的时间序列数据进行分析。

1. 时间序列数据的折线图

表 3.6　某天猫店铺 2016—2019 年季度零售额

序　号	年	季	销售额/万元	序　号	年	季	销售额/万元
1	2016	1	340	9	2018	1	530
2		2	210	10		2	480
3		3	300	11		3	520
4		4	360	12		4	670
5	2017	1	460	13	2019	1	690
6		2	410	14		2	580
7		3	450	15		3	620
8		4	570	16		4	750

单击“插入”图表中的“折线图”选项，如图 3.15 所示。

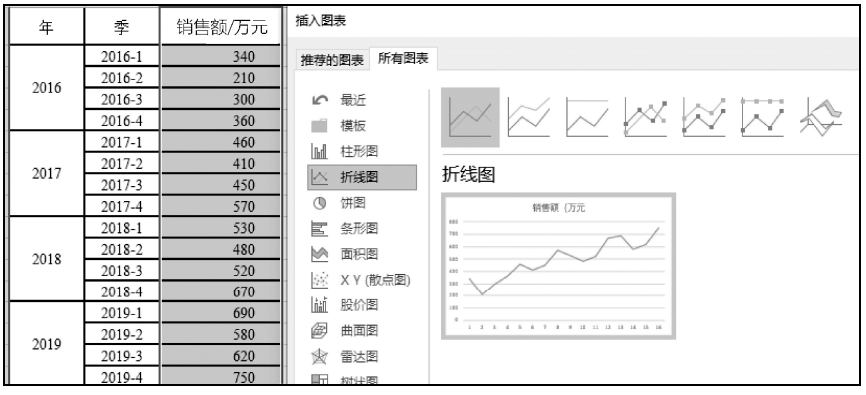


图 3.15　折线图选项

得到销售额时间序列数据的折线图，如图 3.16 所示。

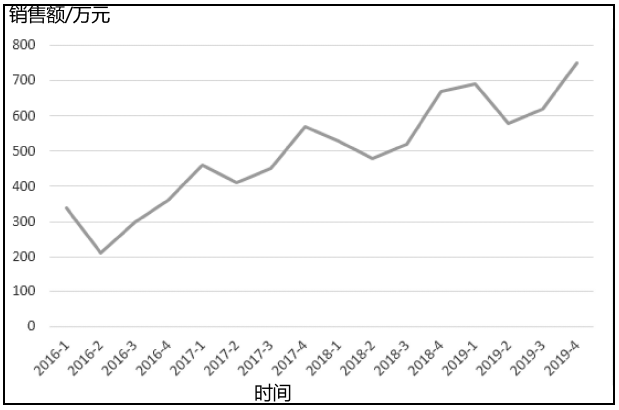


图 3.16　销售额时间序列数据的折线图

2. 时间序列数据的一次移动平均

在数据分析的“分析工具”下拉列表框中选择“移动平均”选项，单击“确定”按钮，如图 3.17 所示。

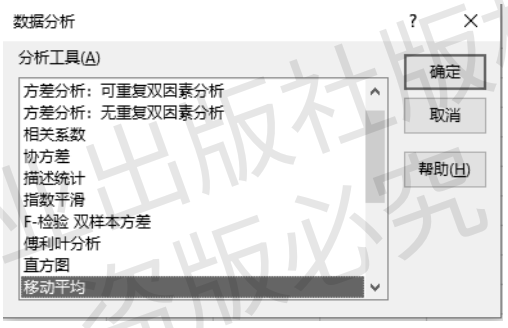


图 3.17　移动平均功能选项

输入区域设置为“$C$1：$C$17”；间隔设置为 3；输出区域设置为“$D$2”，勾选“图表输出”复选框，单击“确定”按钮，如图 3.18 所示。

	A	B	C
1	年	季	销售额/万元
2	2016	2016-1	340
3		2016-2	210
4		2016-3	300
5		2016-4	360
6	2017	2017-1	460
7		2017-2	410
8		2017-3	450
9		2017-4	570
10	2018	2018-1	530
11		2018-2	480
12		2018-3	520
13		2018-4	670
14	2019	2019-1	690
15		2019-2	580
16		2019-3	620
17		2019-4	750

图 3.18　移动平均的操作选项

在单击“确定”按钮后，出现如图 3.19 所示的一次移动平均结果，以及原值与移动平均值的对比图。

年	季	销售额/万元	
2016	2016-1	340	#N/A
	2016-2	210	#N/A
	2016-3	300	283.33
	2016-4	360	290.00
2017	2017-1	460	373.33
	2017-2	410	410.00
	2017-3	450	440.00
	2017-4	570	476.67
2018	2018-1	530	516.67
	2018-2	480	526.67
	2018-3	520	510.00
	2018-4	670	556.67
2019	2019-1	690	626.67
	2019-2	580	646.67
	2019-3	620	630.00
	2019-4	750	650.00

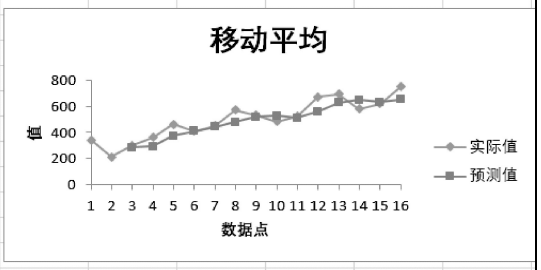


图 3.19　一次移动平均的计算结果

3. 时间序列数据的季节指数趋势法模型

使用如表 3.6 所示的销售额时间序列数据，并采用季节指数趋势法进行计算，如表 3.7 所示。其中，季节比率平均值算法如表 3.8 所示。

表 3.7　各季销售额时间序列数据

年	季	销售额 Y_t/万元	移动平均值 N=4	对正平均值 M_t（N=2）	季节比率 f_t	长期趋势 $\hat{X}_t$	预测值 $\hat{Y}_t$
2016	1	340	—	—	—	288.43	313.8
	2	210	—	—	—	316.14	284.2
	3	300	302.5	317.5	0.9449	343.85	319.2
	4	360	332.5	357.5	1.0070	370.56	402.9
2017	1	460	382.5	401.3	1.1462	399.27	434.5
	2	410	420.0	446.3	0.9187	426.98	383.9
	3	450	472.5	481.3	0.9350	454.69	422.1
	4	570	490.0	498.8	1.1427	482.40	523.1
2018	1	530	507.5	516.3	1.0265	510.11	555.1
	2	480	525.0	537.5	0.8930	537.82	483.6
	3	520	550.0	570.0	0.9123	565.53	525.0
	4	670	590.0	602.5	1.1120	593.24	643.3
2019	1	690	615.0	627.5	1.0996	620.95	675.7
	2	580	640.0	650.0	0.8923	648.66	583.2
	3	620	660.0	—	—	676.37	627.9
	4	750	—	—	—	704.08	763.5

表 3.8　季节比率平均值算法

季	2016	2017	2018	2019	比率合计	平均比率	调整比率
1	—	1.1463	1.0265	1.0996	3.2724	1.0908	1.0881
2	—	0.9187	0.8930	0.8923	2.7040	0.9013	0.8991
3	0.9 449	0.9350	0.9123	—	2.7922	0.9307	0.9284
4	1.0 070	1.1427	1.1120	—	3.2617	1.0872	1.0844

3.4　聚类分析

3.4.1　聚类的定义

聚类（Clustering）是将数据划分成群组的过程，用来研究如何在没有训练的条件下把对象划分为若干类。通过确定数据之间在预先制定的属性上的相似性来完成聚类任务，这样最相似的数据就聚集成簇（Cluster）。聚类与分类不同，聚类的类别取决于数据本身，而分类的类别是由数据分析人员预先定义好的。使用聚类分析算法的用户不但需要深刻地了解所用的特殊技术，而且要知道数据收集过程的细节及拥有应用领域的专家知识。

3.4.2　K-means 算法

K-means 算法接受输入量 k，然后将 n 个数据对象划分为 k 个聚类，以便使获得的聚类满足：同一聚类中数据对象的相似度较高，而不同聚类中数据对象的相似度较小。聚类相似度是利用各聚类中数据对象的均值获得一个“中心对象”（引力中心）来进行计算的。

K-means 算法的工作过程如下。

首先从 n 个数据对象中任意选择 k 个数据对象作为初始聚类中心，而对于剩下的其他数据对象，则根据它们与这些聚类中心的相似度（距离），分别将它们分配给与其最相似的聚类中心代表的聚类；然后计算每个所获新聚类的聚类中心（该聚类中所有对象的均值）。不断重复这一过程直到标准测度函数开始收敛为止。一般采用误差平方和作为标准测度函数，即准则函数 E。

$$E=\sum_{i=1}^{k}\sum_{x\in C_i}d^2(x,z_i) \tag{3-8}$$

设待聚类的数据集为 $X=\{x_1,x_2,\cdots,x_n\}$，将其划分为 k 个簇 C_i，则均值为 z_i，即 z_i 为簇 C_i 的中心(i=1,2,⋯,k)。E 为所有数据对象的误差平方的总和，$x\in X$ 为空间中的点，$d(x,z_i)$为点 x 与 z_i 间的距离，它们可以利用明氏、欧氏、马氏或兰氏距离求得。样本点分类和聚类中心的调整是迭代交替进行的两个过程。

3.4.3　聚类分析算法实例

设有数据样本集合 X={1,5,10,9,26,32,16,21,14}，将 X 聚为 3 类，即 k=3。随机选择前 3 个

数值为初始的聚类中心，即 z_1=1，z_2=5，z_3=10（采用欧氏距离进行计算）。

第一次迭代：按 3 个聚类中心将样本集合分为 3 个簇：{1}，{5}，{10,9,26,32,16,21,14}。对产生的簇分别计算平均值，得到的平均值点填入步骤 2 的 z_1、z_2、z_3 栏中（见表 3.9）。

第二次迭代：通过平均值调整数据对象所在的簇，重新聚类，即将所有点按距离平均值点 1，5，18.3 最近的原则重新分配，得到三个新的簇：{1}，{5,10,9}，{26,32,16,21,14}。将其填入步骤 2 的 C_1、C_2、C_3 栏中（见表 3.9）。重新计算簇平均值点，得到新的平均值点为 1，8，21.8。

以此类推，当进行到第五次迭代时，得到的三个簇与第四次迭代的结果相同，而且准则函数 E 收敛，迭代结束，如表 3.9 所示。

表 3.9　K-means 算法计算过程

步　骤	z_1	z_2	z_3	C_1	C_2	C_3	E
1	1	5	10	{1}	{5}	{10,9,26,32,16,21,14}	433.43
2	1	5	18.3	{1}	{5,10,9}	{26,32,16,21,14}	230.8
3	1	8	21.8	{1}	{5,10,9,14}	{26,32,16,21}	181.76
4	1	9.5	23.8	{1,5}	{10,9,14,16}	{26,32,21}	101.43
5	3	12.3	26.3	{1,5}	{10,9,14,16}	{26,32,21}	101.43

本章知识小结

本章主要学习了与电子商务数据分析相关的模型方法，主要包括两大类，一类是统计分析，包括静态分析指标、动态分析指标、统计指数、抽样推断、相关分析与回归分析等内容；另一类是数据挖掘模型，主要是从大量的数据中发现隐含的、事先未知的、潜在的、有用的信息，或者知识、规则、规律、模式，其主要包括时间序列分析模型、聚类分析算法等。

本章考核检测评价

1. 判断题

（1）总量指标属于静态分析指标。

（2）平均数动态数列是把一系列同类的相对指标数值按时间先后顺序排列形成的。

（3）一般情况下，样本容量超过 20 个样本即可视为大样本。

（4）某次计算得到相关系数 R 为 0.58，可认为两类变量之间存在显著相关。

（5）在指数平滑方法中，若认为近期影响高于远期影响，则加权系数 α 可大一些。

2. 单选题

（1）以下不属于相对指标的是（　　）。

A．倍数　　B．成数　　C．百分数　　D．中位数

（2）按相关程度划分，以下不属于相关关系的是（　　）。

A．完全相关　　B．正相关

C．不相关　　D．不完全相关

（3）求解线性回归模型参数的方法称为（　　）。

A．相关系数　　B．聚类分析

C．最小二乘　　D．关联规则

（4）聚类分析算法停止的标志是（　　）收敛。

A．聚类数　　B．聚类中心

C．准则函数　　D．相似度

（5）能够在没有训练的条件下把对象划分为若干类的算法是（　　）。

A．回归分析　　B．关联规则

C．聚类分析　　D．分类分析

3. 多选题

（1）平均指标包括（　　）。

A．算术平均数　　B．调和平均数

C．众数　　D．中位数

（2）按表示的特征不同，统计指数可分为（　　）。

A．数量指标指数　　B．定基指数

C．质量指标指数　　D．动态指数

（3）以下属于按形式分类的相关关系的有（　　）。

A．线性相关　　B．非线性相关

C．偏相关　　D．曲线相关

（4）以下可以被认为是时间序列数据的有（　　）。

A．股票价格　　B．每日气温

C．心跳变化　　D．地震强度

（5）在聚类分析中，可作为测量类间距离方法的有（　　）。

A．明氏距离　　B．欧氏距离

C．马氏距离　　D．兰氏距离

4. 简答题

（1）什么是变异指标？

（2）简述统计指数的作用。

（3）简述回归分析模型的主要形式。

（4）简述 K-means 算法的工作过程。

（5）简述季节指数水平法的主要步骤。

5. 案例题

（1）Target 百货商店的怀孕预测指数案例。

数据分析可以驱动市场营销、成本控制、产品和服务、管理和决策及商业模式的创新。Target

顾客数据分析部高级经理 Andrew Pole 根据 Target 迎婴聚会（Baby Shower）的登记表建立了一个购买商品与妊娠阶段之间的相关模型，选出了 25 种典型商品的消费数据，构建了“怀孕预测指数”。通过这个指数，Target 能够在很小的误差范围内预测到顾客的怀孕情况，并把孕妇优惠广告寄发给顾客。根据 Andrew Pole 的数据模型，Target 制订了全新的广告营销方案，结果 Target 的孕期用品销售呈现了爆炸性的增长。然后 Target 从孕妇这个细分顾客群向其他各种细分客户群推广。

根据上述案例思考以下问题。

（1）在商业领域，数据分析的作用有哪些？

（2）上述案例中应用了哪些数据分析方法？

（3）数据分析与保护用户隐私之间，应如何注意数据分析的伦理性？

（4）请同学们分组讨论，还有哪些成功的数据分析案例？