第3章 电子商务系统的指标体系

电子商务指标对于了解电子商务企业的经营运行状况和环境有着极其重要的意义。通常电子商务指标体系分为网站运营指标、经营环境指标、销售指标、营销活动指标和客户价值指标5大类。

3.1 网站运营指标

在电子商务网站的运营中，管理者需要及时了解网站的运营状况，因此针对网站的登录、浏览、交易等各类数据进行分析，已经成了每个网站运营和网络营销工程师每天的必备功课。通过统计和分析网站的运营指标可以帮助我们准确地抓住用户动向和网站的实际状况。根据电子商务网站类型的不同以及要了解的问题的差异，可以有许多不同的指标来衡量。通常来说，网站运营指标有流量指标商品类目指标和供应链指标。

3.1.1 流量指标

流量指标又分成数量指标、流量质量指标和流量转化指标。流量数量指标分为页面浏览量(Page View，PV)、访问人数(Unique Visitor，UV)和访问次数(Visits)。网页浏览量是评价网站流量最常用的指标之一，监测网站PV的变化趋势和分析其变化原因是很多站长定期要做的工作。Page View中的Page一般是指普通的HTML网页，也包含PHP、JSP等动态产生的HTML内容。来自浏览器的一次HTML内容请求会被看作一个PV，逐渐累计成为PV总数。UV是指不同的、通过互联网访问、浏览这个网页的自然人数。Visits表示用户访问的次数。例如：某个网站在1个月内一共有3个UV，90个Visits，180个PV，它们表示的含义是，1个月里面，该网站有3个用户，他们共访问网站90人次，在这90次访问中他们共访问了180个网页。通常对于网站来说，大部分的网站都是这样，一个月内大部分用户都只来一次，所以，经常看到的情况是UV和Visits非常接近，但这两者是完全不同的概念。

流量质量指标分为跳出率(Bounce Rate)、页面/网站停留时间和PV与UV比。跳出率，是网站分析中的一个量度，跳出率定义了只浏览单个页面的访问量占总访问量的比例。这里的跳出(Bounce)指成功进入网站(如果不能成功打开网站就关闭，则称为一个Loss)后，不单击页面上的任何链接，就关闭它，那么对该网站而言就是一个跳出。这个指标是所有内容型指标中最重要的一个，通常认为首页是最高的进入页面(当然，如果网站

有其他更高的进入页面，那么也应该把它加入追踪的目标中，如推广广告等）。对任意一个网站，可以想象，如果访问者对首页或最常见的进入页面都是一掠而过的，说明网站内容不够吸引人，网站策划时在某一方面可能有问题。如果针对的目标市场是正确的，说明是访问者不能找到他想要的东西，或者是网页的设计有问题（包括页面布局、网速、链接的文字等）；如果网站设计是可行易用的，网站的内容可以很容易地找到，那么问题可能出在访问者的质量上，即市场问题。页面/网站停留时间，顾名思义，可以理解为一个用户在一个网页或者网站上花费的时间（Time Spent On A Page/Website）。PV 与 UV 比表示了平均每个用户阅读网页的数量，反映了平均每个用户给网站带来的 PV 数。

除此之外，浏览页面比（Scanning Page Ratio），浏览用户比（Scanning Visitor Ratio）和浏览用户指数（Scanning Visitor Index）等指标可用来描述用户给网站带来的流量质量。

浏览页面比的计算公式是浏览用户量＝少于 1 分钟的浏览页数/所有浏览页数，该指标的意义是在一分钟内完成的访问页面数的比例。指标用法：根据网站的目标不同，这个指标的高低有不同的要求，大部分的网站希望这个指标降低。如果是搞广告驱动的网站，这个指标太高对于长期的目标是不利的，因为这意味着尽管通过广告吸引了许多访问者，产生了很高的访问页数，但是访问者的质量却是不高的，所能带来的收益也就会受到影响。

浏览用户比的计算公式是浏览用户比＝少于 1 分钟的访问者数/总访问人数，这个指标在一定程度上衡量了网页的吸引程度。指标用法：大部分的网站都希望访问者停留超过一分钟，如果这个指标的值太高，那么就应该考虑一下网页的内容是否过于简单，网站的导航菜单是否需要改进。该指标与浏览页面比的区别在于针对的对象不同，该指标是描述用户的，而前一个指标描述的对象是页面。

浏览用户指数的计算公式：浏览用户指数＝少于 1 分钟的访问页面数/少于 1 分钟的访问者数，指标意义是一分钟内的访问者平均访问页数。指标用法：这个指数越接近于 1，说明访问者对网站越没兴趣，他们仅仅是瞄一眼就离开了。这也许是导航的问题，如果对导航系统进行了显著的改进，应该可以看到这个指数在上升；如果指数还是下降，应该是网站的目标市场及使用功能有问题，应该着手解决。将浏览用户比和浏览用户指数结合起来使用，可以看出用户是在浏览有用的信息还是厌烦而离开。

流量转化指标分为转化次数和转化率两种指标，转化次数（Conversions），也叫作转化页面到达次数，指独立访客达到转化目标页面的次数。转化率（Conversion Rate）指在一个统计周期内，完成转化行为的次数占推广信息总单击次数的比。计算公式为：转化率＝（转化次数/点击量）×100％。两者是紧密相连的两个概念。例如：10 名用户看到某个搜索推广的结果，其中 5 名用户单击了某一推广结果并被跳转到目标 URL 上，之后，其中 2 名用户有了后续转化的行为。那么，这条推广结果的转化率就是（2/5）×100％＝40％。转化率是网站最终能否盈利的核心，提升网站转化率是网站综合运营实力的结果。其主要目的是衡量网站内容对访问者的吸引程度以及网站的宣传效果。指标用法：

在不同的地方测试新闻订阅、下载链接或注册会员，可以使用不同的链接名称、订阅方式、广告放置、付费搜索链接、付费广告（PPC）等，看看哪种方式能够保持转换率上升？如何增强来访者和网站内容的相关性？如果这个值上升，说明相关性增强了，反之，则是减弱了。

3.1.2 商品类目指标

商品类目指标主要是用来衡量网站商品的正常运营水平，这一类目指标与销售指标以及供应链指标关联紧密。如商品类目结构占比，各类目销售额占比，各类目销售 SKU 集中度以及相应的库存周转率等，不同的产品类目占比又可细分为商品大类目占比情况以及具体商品不同大小、颜色、型号等各个类别的占比情况等。

商品类目结构占比，各个类目商品数量占整体商品数量的比例，体现了商品销售的结构以及商品数量的丰富度和多样性。商品类目销售额占比是指各个类目商品销售额占整体商品销售额的比例。类目销售库存量单位（Stock Keeping Unit，SKU）集中度则表示不同类型、型号和规格的产品集中程度。库存周转率（Inventory Turn Over ，ITO），通常衡量一种材料在工厂里或是整条价值流中的流动快慢。最常见的计算库存周转率的方法，就是把年度销售产品的成本（不计销售的开支以及管理成本）作为分子，除以年度平均库存价值。因此，库存周转率＝年度销售产品成本/当年平均库存价值，该公式对于电子商务企业仍然适用。

3.1.3 供应链指标

这里的供应链指标主要指电商网站商品库存以及商品发送方面，而关于商品的生产以及原材料库存运输等则不在考虑范畴之内。这里主要考虑从顾客下单到收货的时长、仓储成本、仓储生产时长、配送时长、每单配送成本等。譬如仓储中的分仓库压单占比、系统报缺率（与前面的商品类目指标有极大的关联）、实物报缺率、限时上架完成率等，物品发送中的譬如分时段下单出库率、未送达占比以及相关退货比、COD 比等。

3.2 经营环境指标

经营环境指标分为外部竞争环境指标和内部购物环境指标。外部竞争环境指标主要包括网站的市场占有率，市场增长率，网站排名等。网站内部购物环境指标包括功能性指标和运营指标（这部分内容和之前的流量指标是一致的），常用的功能性指标包括商品类目多样性、支付配送方式多样性、网站正常运营情况、链接速度等。

3.2.1 外部竞争指标

外部竞争指标包括市场占有率、市场增长率、网站排名和访问比重等。市场占有率也叫市场份额（Market Shares），是指一个企业的销售量（或销售额）在市场同类产品中所占

的比重，直接反映消费者和用户对企业所提供的商品和劳务的满足程度，表明企业的商品在市场上所处的地位。市场份额是企业的产品在市场上所占的份额，也就是企业对市场的控制能力。市场份额越高，表明企业经营、竞争能力越强。企业市场份额的不断扩大，可以使企业获得某种形式的垄断，这种垄断既能带来垄断利润又能保持一定的竞争优势。这种扩大的趋势可以用市场增长率来表示，市场增长率是指产品或劳务的市场销售量或销售额在比较期内的增长率，其计算公式如下：

市场增长率＝［比较期市场销售量(额)－前期市场销售量(额)］
÷前期市场销售量(额)×100%

一般所指的网站排名主要可分为几大类，有 Alexa 网站排名，中国网站排名，百度，NNT 网站排名等。对于任何一家网站来说，要想在网站推广中取得成功，搜索引擎优化都是最为关键的一项任务。同时，随着搜索引擎不断变换它们的排名算法规则，每次算法上的改变都会让一些排名很好的网站在一夜之间名落孙山，而失去排名的直接后果就是失去了网站固有的可观访问量。所以每次搜索引擎算法的改变都会在网站之中引起不小的骚动和焦虑。可以说，搜索引擎优化是一个越来越复杂的任务。除了这些排名之外，对于电子商务网站来说，还有在一定时期内网站交易额在同类购物网站中的排名和独立访客数在同类购物网站中的排名。

访问比重是对一个站点下属栏目或子站点访问量的统计，较常用的是独立访问人数占同类同期所有网站合计人数的比重。也可以统计一定时期其他流量指标的比重，例如页面浏览量(PV)、访问人数(UV)和访问次数等流量指标在同类网站中的比重。

3.2.2 内部购物指标

内部购物指标包括运营指标和基本功能指标，反映了网站的运营经营状况以及实现的功能。运营指标同样包含了页面浏览量(PV)、访问人数(UV)和访问次数等流量指标，也包含了访问到购物车转化率，访问到下单转化率，下单到支付的订单转化率和订单数量以及金额等。基本功能指标包含了支付方式、配送方式、商品数目和最短流程等方面的指标。

3.3 销售业绩指标

销售业绩指标直接与公司的财务收入挂钩，这一块指标在所有数据分析指标体系中起提纲挈领的作用，其他数据指标的细化落地都可以根据该指标去细分。这里销售业绩指标分解为网站销售业绩指标和订单销售业绩指标，其实两者并没有太大的区别，网站销售业绩指标重点在网站订单的转化率方面，而订单销售指标重点则在具体的毛利率、订单有效率、重复购买率、退换货率方面，当然还有很多指标，譬如总销售额、品牌类目销售额、总订单、有效订单等。

3.3.1 网站销售业绩指标

网站销售业绩指标包括：下单次数，加入购物车次数，在线支付次数，购物车转化率，下单转化率和支付转化率。

下单次数是指在一个统计周期内，购物网站上客户提交订单的次数。

加入购物车次数是指在一个统计周期内，客户单击货物放入购物车和立即购买的次数。

在线支付次数是指在一个统计周期内，购物网站完成购物流程，成功在线支付的次数。

访问到购物车的转化率，在一个统计周期内，购物网站下单次数与访问该网站的次数之比。

下单到在线支付的转化率，在一个统计周期内，购物网站在线支付次数与下单次数之比。

3.3.2 订单指标

订单销售指标包括毛利率、订单有效率、重复购买率、退换货率方面以及总销售额、品牌类目销售额、总订单、有效订单等。

除此之外，与订单指标密切相关的指标还有：平均订货额(Average Order Amount，AOA)，订单转化率(Conversion Rate，CR)，每访问者销售额(Sales Per Visit，SPV)，单笔订单成本(Cost Per Order，CPO)，再订货率(Repeat Order Rate，ROR)等。

平均订货额用来衡量网站销售状况的好坏，其计算公式是平均订货额＝总销售额/总订货数。指标用法：将网站的访问者转化为买家当然是很重要的，同样重要的是激励买家在每次访问时购买更多的产品。跟踪这个指标可以找到更好的改进方法。

订单转化率的计算公式是订单转化率＝总订货数/总访问量，该指标的意义是这是一个比较重要的指标，衡量网站对每个访问者的销售情况。指标用法：通过这个指标可以看到即使一些微小的变化都可能给网站的收入带来巨大的变化。如果还能够区分出新、旧访问者所产生的订单，那么就可以细化这个指标，对新旧客户分别进行统计。

每访问者销售额的计算公式：每访问者销售额＝总销售额/总访问数，指标意义：这个指标也是用来衡量网站的市场效率的。指标用法：这个指标和转化率差不多，只是表现形式不同。

单笔订单成本用来衡量平均的订货成本，其计算公式是单笔订单成本＝总的市场营销开支/总订货数。指标用法：每笔订单的营销成本对于网站的盈利和现金流都是非常关键的。营销成本的计算各人有不同的标准，有些把全年的网站营运费用摊入每月的成本中，有些则不这么做，关键要看哪种最适合自己的情况。如果能够在不增加市场营销成本的情况下提高转化率，这个指标就应该会下降。

再订货率的计算公式：再订货率＝现有客户订单数/总订单数，该指标用来衡量网站

对客户的吸引力。指标用法：这个指标的高低和客户服务有很大的关系，只有满意的用户产品体验和服务才能提高这个指标。

单个访问者成本的计算公式：单个访问者成本＝市场营销费用/总访问数。该指标的意义：用来衡量网站的流量成本。指标用法：这个指标衡量的是市场效率，目标是降低这个指标而提高 SPV，为此要将无效的市场营销费用削减，增加有效的市场投入。

订单获取差额的计算公式：订单获取差额＝单个访问者成本(CPV)－单笔订单成本(CPO)。指标意义：这是一个衡量市场效率的指标，代表着网站所带来的访问者和转化的访问者之间的差异。指标用法：指标的值应是一个负值，这是一个测量从非访问者中获得客户的成本。有两种方法来降低这个差额，增强了网站的销售能力，CPO 就会下降，这个差额就会缩小，说明网站转化现有流量的能力得到了加强；同样地，CPV 可能升高而 CPO 保持不变或降低，这个差额也会缩小，表明网站所吸引的流量都具有较高的转化率，这种情形通常发生在启用了 PPC(Pay Per Click)计划的情况下。

订单获取率的计算公式是订单获取率＝单笔订单成本(CPO)/单个访问者成本(CPV)，该指标的意义是用另一种形式来体现市场效率。指标用法：用比值的形式往往比较容易为管理阶层所理解，尤其是财务人员。

每笔订单产出的计算公式：每笔产出＝(平均订货数×平均边际收益)－每笔订单成本。该指标的意义是每笔订单带来的现金增加净值。指标用法：公司的财务总监总是对这个指标感兴趣的，代表花了多少钱来赚多少钱。

3.4 营销活动指标

一场营销活动做得是否成功，通常从活动效果(收益和影响力)、活动成本以及活动黏合度(通常以用户关注度、活动用户数以及客单价等来衡量)等几方面考虑。营销活动指标区分为日常市场运营活动指标、广告投放指标以及对外合作指标。

市场运营活动指标和广告投放指标主要考虑新增访客数、订单数量、下单转化率、每次访问成本、每次转换收入以及投资回报率(Return on Investment，ROI)等指标。新增访客数、订单数量这两个概念较容易理解，下单转化率是指支付次数/下单次数，也就是转化为最终成交的订单数之比。网上广告收费最科学的办法是按照有多少人看到该广告来收费。按访问人次收费已经成为网络广告的惯例。每次访问成本是指听到或者看到某广告的每人平均分担到多少广告成本。投资回报率的计算公式是投资回报率＝每笔产出(CON)/每笔订单成本(CPO)，用来衡量广告的投资回报。指标用法：比较广告的回报率，应该把钱分配给有最高回报率的广告，但是这个回报率应当要有时间段的限制，如“25% RIO/每周”和“25% RIO/每年”是有很大差别的。这些指标主要反映了营销活动对电子商务网站所带来的积极效果的方方面面，但是这些指标往往是短效的，不一定能够全面评估一次营销活动的效果。

对外合作指标则根据具体合作对象而定，合作的对象可以是其他网站、媒体和机构。

譬如某电商网站与返利网合作，首先考虑的也是合作回报率，可以把合作回报率当作评价合作质量的一个重要指标。

3.5 客户价值指标

一个客户的价值通常由三部分组成：历史价值（过去的消费）、潜在价值（主要从用户行为方面考虑，RFM 模型为主要衡量依据）、附加值（主要从用户忠诚度、口碑推广等方面考虑）。这里客户价值指标分为总体客户指标以及新、老客户价值指标，这些指标主要从客户的贡献和获取成本两方面来衡量。譬如，这里用访客人数、访客获取成本以及从访问到下单的转化率来衡量总体客户价值指标，而对老顾客价值的衡量除了上述考虑因素外，更多的是以 RFM 模型为考虑基准。

3.5.1 客户指标

客户指标包括访问人数，访客获得成本和访问到下单的转化率等重要指标。这里的访问人数就是在一个统计周期内，购物网站的独立访问用户数，也就是前面提到的 UV。访客获得成本是指获得一个新访客所花费的营销、宣传成本之和。访问到下单的转化率就是在一个统计周期内，提交订单的访问数与总访问数之比。

3.5.2 新客户指标

新客户是网站反映网站用户数量发展的一个重要指标。新客户指标包括新顾客数量，新顾客获取成本和客单价。新顾客数量较容易理解，是指在一个统计周期内独立访问网站并产生一次购物的用户数。新客户成本是指企业为吸引客户而花费的各类资源，包括花费在宣传促销、经营、计划、服务以及营销部门的某些销售费用等活动上的费用。新顾客的客单价（Per Customer Transaction）是指网站（超市）每一个新顾客每笔订单的平均交易金额。

3.5.3 老客户指标

这里的老客户是指在一个统计周期内，完成两次或者两次以上购物的总用户数。当新客户回访网站后会变成老客户，维护老客户的活跃度对于电子商务网站同样非常重要，他们的行为会对企业的业绩产生非常重要的影响。这里涉及回访者比（Repeat Visitor Share）这个概念，其计算公式为回访者比＝回访者数/独立访问者数，指标意义是衡量网站内容对访问者的吸引程度和网站的实用性，网站是否有令人感兴趣的内容使访问者再次回到该网站。指标用法：基于访问时长的设定和产生报告的时间段，这个指标可能会有很大的不同。绝大多数网站都希望访问者回访，因此都希望这个值不断提高，如果这个值在下降，说明网站内容或产品的质量没有加强。需要注意的是，一旦选定了一个时长和时间段，就要使用相同的参数来产生报告，否则就失去比较的意义。

老客户指标包含老顾客数量以及 RFM，即最后一次消费时间（Recency），消费频率（Frequency）和消费金额（Monetary）等。

最近一次消费意指上一次购买的时候——顾客上一次是几时来店里、上一次根据哪本邮购目录购买东西、什么时候买的车，或最近一次在超市买早餐是什么时候。理论上，上一次消费时间越近的顾客应该是越好的顾客，对提供即时商品或服务也最有可能会有反应。营销人员若想业绩有所成长，只能靠偷取竞争对手的市场占有率，而如果要密切地注意消费者的购买行为，那么最近一次消费就是营销人员第一个要利用的工具。历史显示，如果能让消费者购买，他们就会持续购买。这也就是为什么，0～6 个月的顾客收到营销人员的沟通信息多于 31～36 个月的顾客。最近一次消费报告是维系顾客的一个重要指标。最近才买你的商品、服务或是光顾你商店的消费者，是最有可能再向你购买东西的顾客。再则，要吸引一个几个月前才上门的顾客购买，比吸引一个一年多以前来过的顾客要容易得多。营销人员如接受这种强有力的营销哲学——与顾客建立长期的关系而不仅是卖东西，会让顾客持续保持往来，并赢得他们的忠诚度。最近一次消费的功能不仅在于提供的促销信息而已，营销人员最近一次的消费报告可以监督事业的健全度。优秀的营销人员会定期查看最近一次消费分析，以掌握趋势。月报告如果显示上一次购买很近的客户，（最近一次消费为 1 个月）人数如增加，则表示该公司是个稳健成长的公司；反之，如上一次消费为一个月的客户越来越少，则是该公司迈向不健全之路的征兆。

消费频率是顾客在限定的期间内所购买的次数。可以说最常购买的顾客，也是满意度最高的顾客。如果相信品牌及商店忠诚度的话，最常购买的消费者，忠诚度也就最高。增加顾客购买的次数意味着从竞争对手处偷取市场占有率，从别人的手中赚取营业额。根据这个指标，又把客户分成五等分，这个五等分分析相当于一个“忠诚度的阶梯”（Loyalty Ladder），其诀窍在于让消费者一直顺着阶梯往上爬，把销售想象成要将两次购买的顾客往上推成三次购买的顾客，把一次购买者变成两次。

消费金额是所有数据库报告的支柱，也可以验证“帕雷托法则”（Pareto's Law）——公司 80％的收入来自 20％的顾客。它显示出排名前 10％的顾客所花费的金额比下一个等级者多出至少 2 倍，占公司所有营业额的 40％以上。如看累计百分比的那一栏，会发现有 40％的顾客贡献公司总营业额的 80％；而有 60％的客户占营业额的 90％以上。最右一栏显示每等分顾客的平均消费，表现最好的 10％的顾客平均花费 1195 美元，而最差的 10％仅有 18 美元 。

案例分析：如果预算不多，而且只能提供服务信息给 2000 或 3000 个顾客，你会将信息邮寄给贡献 40％收入的顾客，还是那些不到 1％的顾客？数据库营销有时候就是这么简单。这样的营销所节省下来的成本会很可观 。

结合这三个指标，就可以把顾客分成 5×5×5 ＝ 125 类，对其进行数据分析，然后制定营销策略。最近一次消费、消费频率、消费金额是测算消费者价值最重要也是最容易的方法，这充分表现了这三个指标对营销活动的指导意义。而其中，最近一次消费是最有力的预测指标。

在众多客户关系管理(CRM)的分析模式中,RFM 模型是被广泛提到的。RFM 模型是衡量客户价值和客户创利能力的重要工具和手段。该模型通过一个客户的近期购买行为、购买的总体频率以及花了多少钱三项指标来描述该客户的价值状况。

RFM 模型较为动态地展示了一个客户的全部轮廓,这对个性化的沟通和服务提供了依据,同时,如果与该客户打交道的时间足够长,也能够较为精确地判断该客户的长期价值(甚至是终身价值),通过改善三项指标的状况,从而为更多的营销决策提供支持。

在 RFM 模式中,R(Recency)表示客户最近一次购买的时间有多远,F(Frequency)表示客户在最近一段时间内购买的次数,M (Monetary)表示客户在最近一段时间内购买的金额。一般的分析型 CRM 着重于对客户贡献度的分析,RFM 则强调以客户的行为来区分客户。

RFM 非常适用于生产多种商品的企业,而且这些商品单价相对不高,如消费品、化妆品、小家电、录像带店、超市等;它也适合在一个企业内只有少数耐久商品,但是该商品中有一部分属于消耗品,如复印机、打印机、汽车维修等消耗品;RFM 对于加油站、旅行保险、运输、快递、快餐店、KTV、行动电话信用卡、证券公司等也很适合。

RFM 可以用来提高客户的交易次数。业界常用的 DM(直接邮寄),常常一次寄发成千上万封邮购清单,其实这是很浪费钱的。根据统计(以一般邮购日用品而言),如果将所有 R(Recency)的客户分为五级,最好的第五级回函率是第四级的三倍,因为这些客户刚完成交易不久,所以会更注意同一公司的产品信息。如果用 M(Monetary)来把客户分为五级,最好与次好的平均回复率,几乎没有显著差异。

有些人会用客户绝对贡献金额来分析客户是否流失,但是绝对金额有时会曲解客户行为。因为每个商品价格可能不同,对不同产品的促销有不同的折扣,所以采用相对的分级(例如 R、F、M 都各分为五级)来比较消费者在级别区间的变动,则更可以显现出相对行为。企业用 R、F 的变化,可以推测客户消费的异动状况,根据客户流失的可能性,列出客户,再从 M(消费金额)的角度来分析,就可以把重点放在贡献度高且流失机会也高的客户上,重点拜访或联系,以最有效的方式挽回更多的商机。

RFM 也不可以用过头,而造成高交易的客户不断收到信函。每一个企业应该设计一个客户接触频率规则,如购买三天或一周内应该发出一个感谢的电话或 E-mail,并主动关心消费者是否有使用方面的问题,一个月后发出使用是否满意的询问,而三个月后则提供交叉销售的建议,并开始注意客户的流失可能性,不断创造主动接触客户的机会。这样一来,客户再购买的机会也会大幅提高。

企业在推行 CRM 时,就要根据 RFM 模型的原理,了解客户差异,并以此为主轴进行企业流程重建,才能创新业绩与利润。否则,将无法在新世纪的市场立足。

除了 RFM 之外还包含积极访问者比、忠实访问者比、忠实访问者指数,忠实访问者量,它们的具体计算公式和含义如下:

(1) 积极访问者比(Heavy User Share)。

计算公式:积极用户比=访问超过 N 页的用户/总访问数

指标意义：衡量有多少访问者是对网站的内容高度感兴趣的。

指标用法：根据网站的内容和大小，去衡量 N 的大小，如内容类的网站通常定义为11～15页，如果是电子商务类网站则可定义为7～10页。如果网站针对正确的目标受众并且网站使用方便，可以看到这个指标应该是不断上升的。

(2) 忠实访问者比(Committed Visitor Share)。

计算公式：访问时间在 N 分钟以上的用户数/总用户数

指标意义：和上一个指标的意义相同，只是使用停留的时间取代浏览页数，取决于网站的目标，可以使用两个中的一个或结合使用。

指标用法：其中的 N 也通过网站的类型和大小来定义，如大型网站通常定位在20分钟左右。这个访问者指标如果单独使用很难体现它的效用，应该结合其他网站运营的数据指标一起使用，例如转换率，但总体来说，较长的访问时长意味着用户喜欢待在该网站，高的忠实访问率当然是较好的。同样地，访问时长也可以根据不同的需要自行设定。

(3) 忠实访问者指数(Committed Visitor Index)。

计算公式：忠实访问者指数＝大于 N 分钟的访问页数/大于 N 分钟的访问者数

指标意义：指的是每个长时间访问者的平均访问页数，这是一个重要的指标，它结合了页数和时间。

指标用法：这个指数通过页面和时间对网站进行了一个更细的区分，也许访问者正好离开吃饭去了。如果这个指数较低，那意味着有较长的访问时间但是较低的访问页面。通常都希望看到这个指数有较高的值，如果修改了网站，增加了网站的功能和资料，吸引更多的忠实访问者留在网站并浏览内容，这个指数就会上升。

(4) 忠实访问者量(Committed Visitor Volume)。

计算公式：忠实访问者量＝大于 N 分钟的访问页数/总访问页数

指标意义：长时间的访问者所访问的页面占所有访问页面数的量。

指标用法：网站通常都是靠宣传和推广吸引用户的，这个指标的意义就显得尤为重要了，因为它代表了总体的页面访问质量。如果有10 000个访问页数却仅有1%的忠实访问者率，这意味着可能吸引了错误的访问者，这些访问者没有价值，他们仅仅看一眼网页就离开了。这时应该考虑推广方式和宣传方式是不是有什么问题。

(5) 访问者参与指数(Visitor Engagement Index)。

计算公式：访问者参与指数＝总访问数/独立访问者数

指标意义：这个指标是每个访问者的平均会话(session)，代表着部分访问者的多次访问趋势。

指标用法：与回访者比不同，这个指标代表着回访者的强烈度，如果有一个非常正确的目标受众不断回访网站，这个指数将大大高于1；如果没有回访者，指数将趋近于1，意味着每一个访问者都有一个新的会话。这个指数的高低取决于网站的目标，大部分内容性和商业性的网站都希望每个访问者在每周/每月有多个会话(session)；但是如客户服务尤其是投诉之类的页面或网站则希望这个指数尽可能接近于1。

课 后 习 题

(1) 电子商务系统的供应链指标有哪些?

(2) 经营环境指标有哪些?

(3) 有哪些方法可以评价客户的潜在价值?

第4章 相关和回归分析

关于相关研究的起因，最早是由法兰西斯·高尔顿(Francis Galton)因量度豌豆的大小，觉察到子代的大小有“均值回归”的现象。1877年他搜集大量人体身高数据后，计算分析高个子父母、矮个子父母以及一高一矮父母的后代各有多少个高个子和矮个子子女，从而把父母高的后代高个子比较多、父母矮的其后代高个子比较少这一定性认识具体化为父母与子女之间在身高方面的定量关系。1888年，高尔登在 *Co-relations and their Measurement, chiefly from metric Data* 一文中，充分论述了“相关”的统计意义，同时正式提出“回归”的概念。

为了研究父代与子代身高的关系，高尔顿对所搜集的1078对父亲及其儿子的身高数据进行分析。他发现这些数据的散点图大致呈直线状态，也就是说，总的趋势是父亲的身高增加时，儿子的身高也倾向于增加。但是，高尔顿对试验数据进行了深入的分析，发现了一个很有趣的现象——回归效应。因为当父亲高于平均身高时，他们的儿子身高比他更高的概率要小于比他更矮的概率；父亲矮于平均身高时，他们的儿子身高比他更矮的概率要小于比他更高的概率。它反映了一个规律，即这两种身高父亲的儿子的身高，有向他们父辈的平均身高回归的趋势。对于这个一般结论的解释是：大自然具有一种约束力，使人类身高的分布相对稳定而不产生两极分化，这就是所谓的回归效应，这是统计学上“回归”的最初含义。

4.1 相关分析

4.1.1 相关分析概念

社会经济现象中，一些现象与另一些现象之间往往存在着依存关系，当用变量来反映这些现象的特征时，便表现为变量之间的依存关系。相关分析(Correlation Analysis)就是对两个或多个变量元素进行分析，研究它们之间是否存在某种相关关系，并对具体有相关关系的现象探讨其相关方向以及相关程度。相关分析是一种常用的用于研究变量之间密切程度的统计方法。

4.1.2 相关分析的种类

根据不同的分类标准，相关关系大体上可以分为以下几种：

(1) 按相关的程度分为完全相关、不完全相关和不相关。

如果一个变量元素的变化完全由另一个变量元素的变化所确定，则称为完全相关。例如，某汽车销售店，所售某一款汽车的售价为 20 万，如果把该款汽车的销售额记为 y，销售车辆数记为 x，则 $y=20x$，销售额的变化完全由销售车辆数决定。

如果两变量彼此互不影响，其变量元素的变化各自独立，称两变量不相关。

如果两变量之间的关系，介于完全相关与不相关之间，称不完全相关。

为了确定相关变量之间的关系，可以根据所收集的数据，例如，每人的身高和体重、某产品的销售额和销售量、一个人的收入水平和其受教育程度等，在直角坐标系上画出散点图。散点图是描述变量之间关系的一种直观方法，从中可以大体看出变量之间的关系形态及关系强度，如图 4.1 所示。

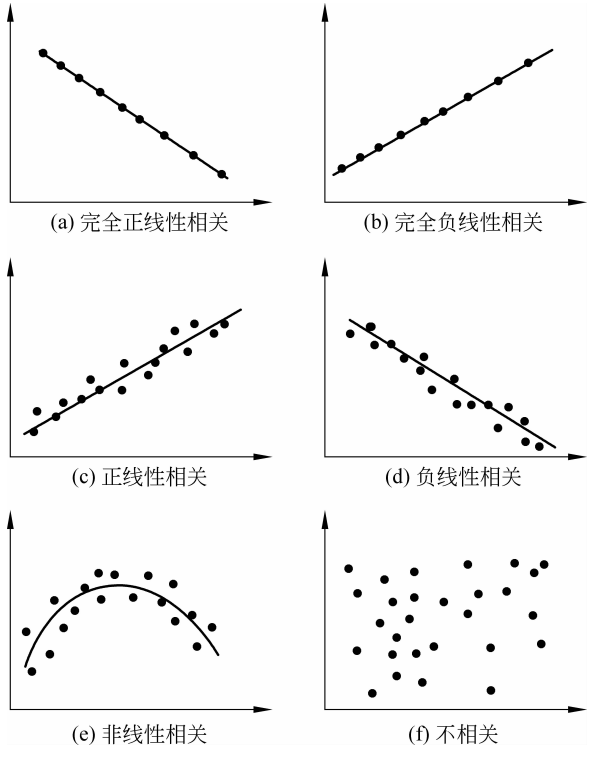

图 4.1　变量间相关的关系

(2) 按变量变化的方向分为正相关和负相关。

若两变量的变动方向一致，即一个变量的数值增加，另一个变量的数值也随之增加，或者一个变量的数值减少，另一个变量的数值也随之减少，则称为正相关。

若两变量的变动方向相反，即一个变量的数值增加，另一个变量的数值也随之减少，或者一个变量的数值减少，另一个变量的数值也随之增加，则称为负相关。

(3) 按相关的形式分为线性相关和非线性相关。

就两个变量而言，如果变量之间的关系近似地表现为一条直线，则称为线性相关；如果变量之间的关系近似地表现为一条曲线，则称为非线性相关或者曲线相关。

(4) 按影响因素的多少分为单相关和复相关。

如果研究的是一个变量同另一个变量之间的相关，就称单相关。

如果分析若干自变量对因变量的影响，称为复相关或多元相关。

4.1.3 相关系数

通过散点图可以帮助判断两个变量之间有无相关关系，并对变量间的关系形态做出大致的描述，但散点图不能准确地反映变量之间的关系强度，为准确地量度两个变量之间的关系强度，还需要计算相关系数。

相关系数是根据样本数据计算两个变量之间的关系强度的统计量。在不同的应用场景下，根据不同的情况，计算相关系数的方法也有很多。

1. 简单相关系数

简单相关系数又叫相关系数或线性相关系数，它是用来量度定量变量间的线性相关关系的。若相关系数是根据总体全部数据计算的，称为总体相关系数，记为 ρ，若是根据样本数据计算的，则称为样本相关系数，一般用字母 r 表示。根据变量类型的不同，主要有三类样本相关系数。设随机变量(X,Y)的 n 对样本$(x_i,y_i)(i=1,2,\cdots,n)$，则三类相关系数分别表示如下：

1) Pearson 相关系数

Pearson 相关也称为积差相关(或积矩相关)，是英国统计学家皮尔逊于 1896 年提出的一种计算直线相关的方法。假设有两个变量 X,Y，那么两变量间的皮尔逊相关系数可通过以下公式计算：

$$r=\frac{\sum_{i=1}^{n}(x_i-\overline{X})(y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{X})^2\sum_{i=1}^{n}(y_i-\overline{Y})^2}}$$

也可以表示为：

$$r=\frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-[E(X)]^2}\ \sqrt{E(Y^2)-[E(Y)]^2}}$$

其中 E 是数学期望。

从上式可以看出，只有当两个变量的标准差都不为零时，公式才有意义，Pearson 相关系数主要适用于：两个变量之间是线性关系，都是连续数据；两个变量的总体是正态分布，或接近正态的单峰分布；两个变量的观测值是成对的，每对观测值之间相互独立；样本中存在的极端值对积差相关系数的影响极大，因此一般要求大样本，并且对极端值要慎重考虑和处理，以避免出现错误的结论。

例 4.1 小学语文教研组为研究四年级学生语文成绩与英语成绩之间的相关程度，从四年级中随机抽取 10 名学生的语文测验成绩和英语测验成绩，如表 4.1 的左半部分所示，试求它们之间的相关程度。

表 4.1 测试学生成绩

学生	语文 x_i	英语 y_i	学生	语文 x_i	英语 y_i
1	85	67	6	77	78
2	81	86	7	87	92
3	84	86	8	90	94
4	78	71	9	78	66
5	72	74	10	87	77

解：用 Excel【数据分析】中的【相关系数】工具计算可得两者的相关系数为：0.56337。

2) Spearman Rank 相关系数

又称秩相关(Rank Correlation)或等级相关，等级相关是指以等级次序排列或以等级次序表示的变量之间的相关。等级相关不受变量总体分布形态的限制，适用于某些不能准确地测量指标值而只能以严重程度、名次先后、反应大小等定出的等级资料，也适用于某些不呈正态分布或难以判断分布的资料。

设 R_i 和 Q_i 分别为 x_i 和 y_i 各自在变量 X 和变量 Y 中的秩，如果变量 X 与变量 Y 之间存在着正相关，那么 X 与 Y 应当是同时增加或减少的，这种现象当然会反映在 (x_i, y_i) 相应的秩 (R_i, Q_i) 上。反之，若 (R_i, Q_i) 具有同步性，那么 (x_i, y_i) 的变化也具有同步性。因此，

$$d = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (R_i - Q_i)^2 \tag{4.1}$$

具有较小的数值。如果变量 X 与变量 Y 之间存在着负相关，那么 X 与 Y 中一个增加时，另一个在减小，d 具有较大的数值。既然由 (x_i, y_i) 构成的样本相关系数反映了 X 与 Y 之间相关与否的信息，那么在参数相关系数的公式 $r(X,Y)$ 中以 R_i 和 Q_i 分别代替 x_i 和 y_i，不是同样地反映了这种信息吗？基于这种想法，Charles Spearman 秩相关系数 $r_s(R,Q)$ 应运而生。

$$r_s(R,Q) = \frac{\sum \left(R_i - \frac{1}{n}\sum R_i\right)\left(Q_i - \frac{1}{n}\sum Q_i\right)}{\sqrt{\sum \left(R_i - \frac{1}{n}\sum R_i\right)^2 \sum \left(Q_i - \frac{1}{n}\sum Q_i\right)^2}} \quad (i = 1,2,\cdots,n) \tag{4.2}$$

$r_s(R,Q)$ 与 $r(X,Y)$ 在形式上完全一致，但在 $r_s(R,Q)$ 中的秩，不管 X 与 Y 取值如何，总是只取 1～n 的数值，因此它不涉及 X 与 Y 总体其他的内在性质。由于

$$\sum_{i=1}^{n} R_i = \sum_{i=1}^{n} Q_i = 1+2+\cdots+n = \frac{n(n+1)}{2}$$

$$\sum_{i=1}^{n} R_i^2 = \sum_{i=1}^{n} Q_i^2 = 1^2+2^2+\cdots+n^2 = \frac{n(n+1)(2n+1)}{6}$$

因此，式(4.2)可以化简为

$$r_s = 1-\frac{6\sum (R_i-Q_i)^2}{n(n^2-1)} = 1-\frac{6\sum d_i^2}{n(n^2-1)} \quad (i=1,2,\cdots,n) \tag{4.3}$$

显然在 $R_i=Q_i$ 时，秩相关系数 r_s 达到最大值1，另外，证明可得秩相关系数 r_s 的最小值为-1。注意：当 X 与 Y 之间相同秩序较多时，应计算 r_s 的校正值。

例 4.2 某公司想要知道是否职工期望成为好的销售员而实际上就能有好的销售记录。为了调查这个问题，公司的副总裁仔细地查看和评价了公司10个职工的初始面试摘要、学科成绩、推荐信等材料，最后副总裁根据他们成功的潜能给出了单独的等级评分。两年后获得了实际的销售记录，得到了第二份等级评分，见表4.2中的第1～第4列。统计问题为是否职工的销售潜能与开始两年的实际销售成绩一致。

表 4.2 职工的销售潜能与销售成绩的秩相关分析

职工编号	潜能等级 R_i	销售成绩	成绩等级 Q_i	$d_i=R_i-Q_i$	d_i^2
1	2	400	1	1	1
2	4	360	3	1	1
3	7	300	5	2	4
4	1	295	6	−5	25
5	6	280	7	−1	1
6	3	350	4	−1	1
7	10	200	10	0	0
8	9	260	8	1	1
9	8	220	9	−1	1
10	5	385	2	3	9
$\sum_{i=1}^{10} d_i^2 =$					44

Spearman 秩相关系数 $r_s(R,Q)$ 的计算过程见表4.2中的第5和第6列，最后计算结果为

$$r_s = 1-\frac{6\sum_{i=1}^{10} d_i^2}{n(n^2-1)} = 1-\frac{6\times 44}{10\times(100-1)} = 0.7333$$

表明潜能与成绩之间有较强的正相关性，高的潜能趋向于好的成绩。

秩相关系数 $r_s(R,Q)$原假设为 0 的 t 检验统计量为

$$t = 0.7333\sqrt{\frac{10-2}{1-(0.7333)^2}} = 3.05$$

自由度为 8，查表 $t=3.05$ 的双侧 $p=0.0158$。在 0.05 显著水平上，t 分布的上临界点为 2.30，由于 $3.05>2.30$，因此，拒绝秩相关系数为 0 的原假设，接受潜能与成绩之间存在秩相关。

2. 复相关系数

复相关系数是反映一个因变量与多个自变量(两个或两个以上)之间相关程度的指标，实际上是描述一个因变量与多个自变量线性组合间的相关关系，复相关系数越大，表明因变量与自变量之间的线性相关程度越密切。例如，某种商品的需求量与其价格水平、职工收入水平等现象之间呈现复相关关系。

为了测定一个变量 y 与其他多个变量 $X_1, X_2, \cdots, X_k$ 之间的复相关系数，可以考虑构造一个关于 $X_1, X_2, \cdots, X_k$ 的线性组合，通过计算该线性组合与 y 之间的简单相关系数作为变量 y 与 $X_1, X_2, \cdots, X_k$ 之间的复相关系数，具体计算过程如下：

第一步，用 y 对 $X_1, X_2, \cdots, X_k$ 作回归，得

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_K \hat{X}_k$$

第二步，计算 y 和$\hat{y}$的简单相关系数，此简单相关系数即为 y 与 $X_1, X_2, \cdots, X_k$ 之间的复相关系数。复相关系数的计算公式为

$$R = \frac{\sum(y_i - \bar{y})(\hat{y} - \bar{y})}{\sqrt{\sum(y-\bar{y})^2 \sum(\hat{y}-\bar{y})^2}}$$

3. 偏相关系数

偏相关系数反映校正其他变量后某一变量与另一变量的相关关系，又叫部分相关系数，是衡量多个变量中某两个变量之间的线性相关程度的指标。假设需要计算 X 和 Y 之间的相关性，Z 代表其他所有变量，X 和 Y 的偏相关系数可以认为是 X 和 Z 线性回归得到的残差 R_x 与 Y 和 Z 线性回归得到的残差 R_y 之间的简单相关系数，即 Pearson 相关系数。

偏相关系数不同于简单相关系数。在计算偏相关系数时，需要掌握多个变量的数据，一方面考虑多个变量之间可能产生的影响，另一方面又采用一定的方法控制其他变量，专门考察两个特定变量的净相关关系。在多变量相关的场合，由于变量之间存在错综复杂的关系，因此偏相关系数与简单相关系数在数值上可能相差很大，有时甚至符号都可能相反。简单相关系数受其他因素的影响，反映的往往是表面的非本质联系。而偏相关系数排除了其他变量的影响，因此用来描述某些变量间的关系更为合理可靠。例如，一种商品的需求既受收入水平的影响又受其价格的影响。按照经济学理论，在一定收入水平下，该商品的价格越高，商品的需求量就越小。也就是说，需求与价格之间应当是负相关的。可

是，在现实经济生活中，由于收入和价格常常都有不断提高的趋势，如果不考虑收入对需求的影响，仅仅利用需求和价格的时间序列数据去计算简单相关系数，就有可能得出价格越高需求越大的错误结论。

在相关分析中，切不可只根据相关系数很大，就认为两个经济变量之间有内在的线性联系或因果关系。因为相关系数只表明两个变量的共变联系，尽管这种共变联系有时也体现了两个变量的内在联系（如物价与需求量），但在很多情况下，这种共变联系是由某个或某些变量的变化所引起的。所以，在研究经济变量之间的相关关系时，当由样本计算的两个变量的相关系数很大时，要认真检查一下这种相关是否与经济理论和经济意义相符，如果不符，一定是由于其他变量的变化所引起的。这时，就需要研究和探索引起这两个变量高度相关的变量，去掉这些变量变化的影响因素，计算偏相关系数，最后确定这两个变量之间的内在线性联系。特别是对时间序列经济变量，一定要考虑去掉时间因素的偏相关系数，否则就会导致荒谬的结论。当研究多个经济变量时，有时计算其中两个变量的相关系数与经济理论和经济意义相符，但由于其他变量影响的作用，这个相关系数可能扩大或缩小了这两个变量之间的真实联系，这时，通过偏相关系数与相关系数的比较，来确定这两个变量之间的内在线性联系会更真实，更可靠。所以，在相关分析中，除了使用相关系数以外，还应该使用偏相关系数，这是非常重要，也是十分必要的。

4. 典型相关系数

在对经济问题的研究和管理研究中，不仅经常需要考察两个变量之间的相关程度，而且还经常需要考察多个变量与多个变量之间即两组变量之间的相关性。典型相关分析就是测度两组变量之间相关程度的一种多元统计方法。典型相关系数量度了两组变量之间联系的强度。对于变量组$(X_1, X_2, \cdots, X_p)$和$(Y_1, Y_2, \cdots, Y_q)$，虽然每个 X_i 与每个 Y_j 之间的相关关系也反映了两组变量中各对之间的联系，但不能反映这两组变量整体之间的相关性。可以把两组变量的相关性转化为两个变量的相关性来考虑，即考察一组变量的线性组合。典型相关分析的基本思想与主成分分析非常相似，先对原来各组变量进行主成分分析，得到新的线性无关的综合指标。再用两组之间的综合指标的直线相关系数来研究原两组变量间的相关关系。首先，在每组变量中找出变量的一个线性组合，例如：

$$U = a_1 X_1 + a_2 X_2 + \cdots + a_p X_p = aX$$

$$V = b_1 Y_1 + b_2 Y_2 + \cdots + b_q Y_q = bY$$

选择 a, b 使得 U, V 之间具有最大的相关系数，据此找出的两个线性组合的变量 U, V 称为第一对典型变量，如果只有一对典型变量还不足以反映 X 和 Y 之间的相关性，可进一步构造与 U, V 互不相关的另外一对典型变量，如此继续下去，还可以确定第三对、第四对典型变量等，并使各对典型变量之间互不相关，直到两组变量之间的相关性被提取完毕为止。这样就将两组变量间的相关性凝结为少数几个典型变量对之间的相关性。称第 k 对典型变量间的相关系数为第 k 个典型相关系数。

典型相关分析计算步骤如下。

(1) 根据分析目的建立原始矩阵。

原始数据矩阵

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} & y_{11} & y_{12} & \cdots & y_{1q} \\ x_{21} & x_{22} & \cdots & x_{2p} & y_{21} & y_{22} & \cdots & y_{2q} \\ & & \vdots & & & & \vdots & \\ x_{n1} & x_{n2} & \cdots & x_{np} & y_{n1} & y_{n2} & \cdots & y_{nq} \end{bmatrix}$$

(2) 对原始数据进行标准化变化并计算相关系数矩阵。

$$\boldsymbol{R} = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$$

其中 R_{11}，R_{22} 分别为第一组变量的相关系数阵和第二组变量的相关系数阵，$R_{12} = R'_{21}$ 为第一组变量与第二组变量的相关系数。

(3) 求典型相关系数和典型变量。

计算矩阵 $\boldsymbol{A} = R_{11}^{-1} R_{12} R_{22}^{-1} R_{21}$ 以及矩阵 $\boldsymbol{B} = R_{22}^{-1} R_{21} R_{11}^{-1} R_{12}$ 的特征值和特征向量，分别得典型相关系数和典型变量。

(4) 检验各典型相关系数的显著性。

首先检验第一对典型变量的相关系数，即

$$H_0 : \hat{\lambda}_1 = 0, \quad H_1 : \hat{\lambda}_1 \neq 0$$

它的似然比统计量为

$$\Lambda_1 = (1 - \hat{\lambda}_1^2)(1 - \hat{\lambda}_2^2) \cdots (1 - \hat{\lambda}_p^2) = \prod_{i=1}^{p} (1 - \hat{\lambda}_i^2)$$

则统计量

$$Q = -\left[n - 2 - \frac{1}{2}(p + q + 1) \right] \ln \Lambda_1$$

给定显著水平 α，查表得 χ_a^2，若 $Q_1 > \chi_a^2$，则否定 H_0，认为第一对典型变量相关，否则不相关，如果相关则依次逐个检验其余典型相关系数，直到某一个相关系数 $\hat{\lambda}_k$ ($k = 2, \cdots, p$) 检验为不显著时截止。

4.2 一元线性回归

相关分析的目的在于测度变量之间的关系强度，它所使用的测度工具就是相关系数。而回归分析则侧重于考察变量之间的数量伴随关系，并通过一定的数学表达式将这种关系描述出来，主要研究一个变量(因变量)关于另外一个或几个变量(自变量)的具体依赖关系。回归分析主要解决以下几个问题：

(1) 从一组样本数据出发，确定变量之间的数学关系式，得到回归方程。

(2) 对回归方程、参数估计值进行各种统计检验，并从影响某一特定变量的诸多变量

中找出哪些变量的影响是显著的，哪些是不显著的。

（3）利用回归方程，根据一个或几个变量的取值来估计或预测另一个特定变量的取值，并给出这种估计或预测的可靠程度。

4.2.1 一元线性回归模型

进行回归分析时，首先需要确定哪个变量是因变量，哪个变量是自变量。在回归分析中，被预测或被解释的变量称为因变量，用 y 表示。用来预测或解释因变量的一个或多个变量称为自变量，用 x 表示。例如，如果研究每月家庭消费支出与每月可支配收入之间的关系，即可根据每月家庭可支配收入，来预测每月家庭的消费支出，则每月家庭可支配收入为自变量 x，每月家庭消费支出为因变量 y。

当回归分析中只涉及一个自变量时称为一元回归，若因变量 y 与自变量 x 之间为线性关系时称为一元线性回归。对于具有线性关系的两个变量，可以用一个线性方程来表示它们之间的关系。描述因变量 y 如何依赖于自变量 x 和误差项 ε 的方程称为回归模型。一元线性回归模型可以表示为

$$y = \beta_0 + \beta_1 x + \varepsilon$$

在一元线性回归模型中，y 是 x 的线性函数（$\beta_0 + \beta_1 x$，其中 β_0，β_1 为模型参数）加上误差项 ε。$\beta_0 + \beta_1 x$ 反映了由于 x 的变化而引起的 y 的线性变化，ε 是被称为误差项的随机变量，反映除了 x 和 y 之间的线性关系之外的随机因素对 y 的影响，是不能由 x 与 y 之间的线性关系所解释的变异性。

描述因变量 y 的期望值如何依赖于自变量 x 的方程称为回归方程，一元线性回归方程的形式为

$$E(y) = \beta_0 + \beta_1 x$$

如果回归方程中的 β_0，β_1 已知，对于一个给定的 x 值，利用上式可计算出 y 的期望值。但总体回归参数 β_0，β_1 是未知的，必须利用样本数据去估算它们。用样本统计量 $\hat{\beta}_0$，$\hat{\beta}_1$ 代替回归方程中的未知参数 β_0，β_1 即可得到估计的回归方程，对一元线性回归，估计的回归方程形式为

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

4.2.2 参数的最小二乘估计

对 x 和 y 的观测值，用于描述其关系的直线有很多，统计学中通常使用最小二乘法，也称最小平方，它是通过使因变量的观测值 y_i 与估计值 $\hat{y}_i$ 之间的离差平方和达到最小来估计 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的方法。最小二乘的思想示意图如图 4.2 所示。

根据最小二乘法使

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$