第 3 章

应用回归分析

回顾第2章，不难发现，类似于高尔夫球手仅仅关心把球打好一样，回归分析仅仅是机械地将数据样本应用于一系列方程中。然而，真正的高尔夫球手会说，如果选错了球杆，或是把球打向沙坑、树木或池塘，都很难将球打好。与此类似，训练有素的计量经济学家考虑其他因素所花费的时间要比单纯考虑回归方程的普通最小二乘法所花费的时间多得多。本章就介绍一些“现实世界”关注的因素。

首先将概述回归分析常用的6个步骤，这是本章最重要的内容。如果读者对其中某个专题（如普通最小二乘法）在回归分析整体框架中所扮演的角色有清晰的认识，那么学习和理解这个专题的能力就会得到提升。另外，在计量经济学研究中，这6个步骤对发展成熟的函数理论起着关键作用。

接下来，介绍一个完整实例。应用Woody’s连锁餐厅的实际数据，通过对连锁店的选址分析，阐述在回归分析中如何应用这6个步骤。在以后的章节，我们将把新的思想和检验应用于这个案例。

3.1 回归分析的步骤

虽然没有硬性规定如何进行计量经济学研究，但多数研究者在回归分析过程中，通常沿用一种标准的方法。虽然每一个步骤的侧重点及其所耗精力有所不同，但所有步骤对一个成功的研究而言都是必要的。值得注意的是，这里并不讨论如何选择被解释变量，它是由研究目的决定的，这方面的内容将在第11章讨论。然而，一旦确定了被解释变量，下面的步骤便顺理成章。

（1）查阅文献，建立理论模型。
（2）确定模型——选择解释变量和函数形式。
（3）对参数的符号做出假设。
（4）搜集、检查和整理数据。
（5）估计和评价方程。
（6）报告结果。

使用这些步骤的目的，并不是不鼓励采用创新的或者非常规的方法，而在于让读者知道：在通常情况下，专业的经济学家和商业分析师如何进行回归分析。

3.1.1 步骤1：查阅文献，建立理论模型

在任何研究中，首要的步骤都是正确把握所研究的课题。通常，最好的数据分析往往是从理论而不是从数据开始的。采用哪些变量以及哪种方程，大多数计量经济学家都是根据基本理论决定的。事实上，若没有切实了解研究的课题，就不可能建立好的计量经济模型。

对绝大多数课题而言，在研究之前，明智的做法是查阅大量相关文献。如果曾经有教授研究过这方面的课题，你就会有兴趣进行下去。如果你的方程中的被解释变量有研究者做过估计，你可能就会采用其中一种模型来处理。另一方面，如果不认同前人的方法，你就可能转向新的方向。无论如何，都不该“另起炉灶”，而应该从前人停下的地方开始。任何关于实证课题的学术论文都应该是基于以前研究的高度总结。

查阅文献最便捷的方法是搜集一些最近出版的《经济文献杂志》或面向企业的出版摘要，还可以开展与课题相关的网络调查或经济学文献检索，如EconLit搜索。利用这些资源，查阅一些近期与课题相关的论文，并注意这些论文的题目。如果一篇早期论文近期被很多作者引用，或者它的标题和你开头的主题吻合，最好追溯并找到这篇文献。在第11章的学习中，我们将会讲述更多有关查阅文献的建议。

有时，一个新的课题可能鲜为人知，因此，无法找到任何相关文献。怎么办呢？在此提供两种可行的方法：一是尝试用相似的理论代替。例如，如果要建立一种新产品的需求模型，那么，可以阅读那些现有同类产品需求的文献。二是向相关领域的工作人士（电话）求助，进行调查。例如，当要建立一个不熟悉的城市住房模型时，请向在当地工作的房地产经纪人函电咨询。

3.1.2 步骤2：确定模型——选择解释变量和函数形式

在应用回归分析中，最重要的就是设定理论回归模型。选定被解释变量后，模型的设定应包含以下3个部分。

（1）解释变量以及如何测量解释变量。

（2）变量间的函数形式。

（3）随机误差项的性质。

若以上部分都处理恰当，回归方程就确定了。我们将分别在第4章、第6章和第7章详细讨论模型设定的细节。

模型设定的每个部分都必须以经济理论为基础。任何一个部分出错都将导致设定误差（specification error）。一般情况下，在所有应用回归分析的错误中，设定误差最糟糕，它将严重影响估计方程的有效性。因此，在课题开始前，越关注经济理论，回归结果就越有保障。

本书重点在于估计行为方程，即描述经济实体行为的方程。着重在经济理论基础上，选择描述有关行为的解释变量。之所以选择解释变量，是因为该解释变量是被解释变量的理论决定因素，并至少可以部分解释被解释变量的变动。请牢记，回归分析只能提供证据而不能证明经济中的因果关系。正如一个例子不能证明规则一样，一个回归结果并不能证明这个理论。

事实上，错误选择解释变量的可能性是存在的。我们的目的仅仅是确定相关的解释变量，在理论上，这些解释变量会对被解释变量产生实质性影响，而不考虑那些对被解释变量几乎没有影响的变量，除非它们对被解释变量可能产生比较特殊（比如政策方面）的潜在影响。

例如，解释某种消费品需求数量的方程也许会将产品价格、消费者收入或财富作为可能的解释变量。在理论上讲，互补品和替代品对需求量也有重要影响。因此，你也许决定把互补品

和替代品的价格考虑进去。那么，应该考虑哪些互补品和替代品呢？当然，考虑相近的互补品和替代品是恰当的，但究竟应该选多少呢？选择和判断必须基于理论上的合理性，而这个过程非常主观。

例如，研究者决定仅仅考虑两种其他产品的价格，认为这两种产品的价格会影响回归方程的前提条件（比如受前期理论支持）或研究假设。前提条件发生变化会使回归方程的待检验假设的数量和类别发生变化。然而，前提条件可能是错误的，由此，就会降低回归分析的有效性。因此，有必要详细说明每个前提条件。

有些定性状态（如性别）由于它们的特征不能被量化，因而无法引入方程。但这些状态可以用虚拟变量（或二元变量）进行量化。虚拟变量（dummy variable）是依据特定状态条件是否成立取值为1或0的变量。

下面，举例来解释虚拟变量。假设：Y_i 表示第i个中学教师的收入，其收入主要依据教师教学经验和获得学位类型而定。虽然所有教师都有学士学位，但有些还具有研究生学位，如硕士学位。于是，表示收入和学位类型两者关系的方程为：

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \tag{3-1}$$

式中，$X_{1i} = \begin{cases} 1 & \text{第}i\text{个教师具有研究生学位；} \\ 0 & \text{其他} \end{cases}$ X_{2i}代表第i个教师的教龄。

变量X_1的取值仅为0或1，所以，X_1被称为虚拟变量，或者仅仅是个“虚拟值”。在这种情况下，虚拟变量表示具有硕士学位。参数β_1是指教师在保持教学经验不变的情况下，拥有更高学位的额外收入。在第7章和第13章中，将详细讨论虚拟变量。

3.1.3 步骤3：对参数的符号做出假设

一旦确定了解释变量，确定回归参数的预期符号就很重要。例如，在最终消费品的需求方程中，需求数量（Q_d）被认为与该消费品的价格（P）以及其互补品的价格（P_c）负相关，与消费者收入（Y）以及其替代品的价格（P_s）正相关。通常，构建回归模型的第1步是建立抽象函数：

$$Q_d = f(\bar{P}, \overset{+}{Y}, \bar{P_C}, \overset{+}{P_s},) + \varepsilon \tag{3-2}$$

式中，变量上面的符号表示线性模型中各个回归参数的预期符号。

许多情况下，基础理论是众所周知的，因此，没有必要讨论每个变量的符号。然而，如果变量的预期符号是不确定的，就应该给出假设参数符号的理由。

3.1.4 步骤4：搜集、检查和整理数据

获得初始数据集，为回归分析做适当的准备不是一件容易的事情。这个步骤不仅要机械地记录数据，还要选择样本的类型和规模。

一般来说，只要观测值都来自同一总体，样本容量越大越好。研究者通常会选用容易获得、有可比性的观测值。在回归分析中，所有变量观测值的数量必须相同，具有相同的频数（月、季、年等）和样本区间。频数的选择往往由获得数据的难易程度决定。

必须选择尽可能多的样本观测值与统计学中的自由度（2.4节中首次提出）概念有关。如图3-1所示，在一个X，Y坐标系中，若过两点作一直线，则可以准确无误地完成。由于两点在同一直线上，因此，不用估计参数，两点决定的两个参数恰好就是截距和斜率。只有用直线拟合由特定程序生成的三个或三个以上不规则点的时候，参数估计才能派上用场。观测值的个数减

去待估计参数的个数（在例子中为2个，即截距和斜率）就是自由度[30]（degrees of freedom）。在图3-2中，需要估计的直线的自由度仅仅为1，而自由度越大越好。因为当自由度较高时，正的误差会被负的误差尽可能地抵消。当自由度较低时，随机因素无法提供能够相互抵消的观测值。例如，抛硬币的次数越多，出现正面的可能性就越接近于真实的概率0.5。

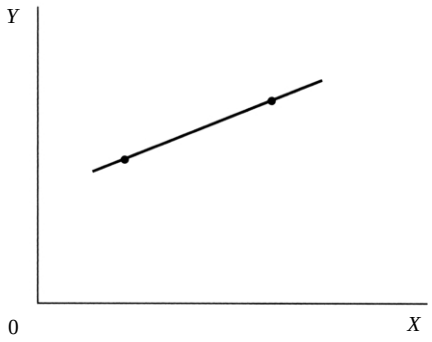

图3-1 数学上两点确定一条直线

注：如图3-1所示，如果数据集中只有两个点，从数学意义上一条直线可以完全无误差地拟合这些点，因为两点决定一条直线。

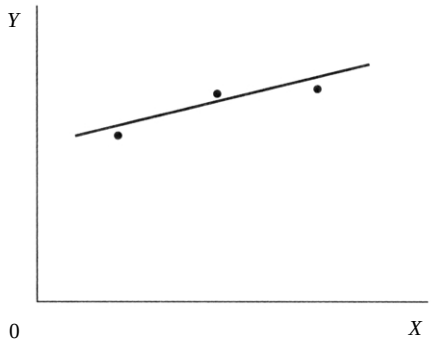

图3-2 统计上直线拟合三个点

注：如图3-2所示，如果数据集中有三个（或更多）点，那么利用2.1节中的估计步骤确定的直线在统计意义下是拟合这些点的。

另一个值得关注的问题是变量的度量单位。某个变量以美元度量或以1 000美元度量有什么区别呢？变量的观测值与真实值相差10个单位会有什么影响呢？有趣的是，就回归分析而言，这些变化并没有什么影响，所有的符号、显著性以及经济理论的结论都与度量单位无关，度量单位仅仅影响参数估计值的大小。例如，一个解释变量的单位是美元或是1 000美元，几乎没什么不同，常数项和模型的整体拟合优度都不变。诚然，变量度量单位的放大或缩小改变了斜率参数，但斜率参数会随着解释变量度量单位的变化而变化。同理，常数被加入到某一变量时，只改变截距项而不会改变斜率参数。

估计方程前的最后一步是检查和整理数据。研究者应该时常检查数据，以便找到错误，原因很简单，如果数据是错的，就没有必要劳神进行复杂的回归分析。

在检查数据的时候，可以将数据打印出来或生成图表以便寻找异常值。异常值（outlier）是指在其余观测值范围外的观测值。找出异常值是找到出错数据的简单方法。另外，查看各个变量的均值、最大值和最小值，并思考数据中可能存在的不一致性也十分必要。有些数据是否不可能存在，或者不切实际？GDP是否可以在一年内翻番？学生是否可能在绩点满分为4.0的条件下，拥有7.0的平均绩点？消费是不是消极的？

一般而言，可以用正确数据替代错误数据来消除这些错误。在极少情况下，诸如无法找到正确数据，或特殊的观测值根本不是和其他样本观测值来自相同总体时，这个异常值才可以从样本中剔除。值得注意的是，异常值的存在并不是剔除该观测值的正当理由。一个回归分析应该可以用来解释样本中所有的观测值，而不仅仅是情况好的数值。关于数据搜集的更多细节，在11.2节和11.3节都有所论述，而有关通过经济学实验形成数据的更多细节会在16.1节讲述。

---

30 通过学习本章，学生将学会计算回归方程中的自由度（d.f.），即$d.f.=(N-K-1)$，其中，K为方程中解释变量的个数。同样，一些作者会设$K'=K+1$并定义$d.f.=(N-K')$。由于K'等于解释变量个数加1（即常数项），故而它也等于在回归分析中要估计的参数个数。

3.1.5 步骤5：估计和评价方程

信与不信，手工完成回归方程的第1步到第4步可能会花费数月时间，但是像EViews 或Stata这样的计算机软件却能在1秒内估计出回归方程。正如2.1节所述，通常情况下，采用普通最小二乘法进行估计。但如果使用其他估计方法，就应该仔细说明和评价选择这种方法的理由。

也许某些人认为估计出方程后，工作就结束了。事实并非如此，相反，还需要用各种不同的方法评价结果。诸如方程拟合程度如何？估计出的参数符号和大小是不是与预期吻合？本书后面的大部分内容讲述的是如何评价估计出的回归方程。对于初学者而言，应该花费较多时间来评价方程。

一旦完成评价，不要自动进入第6步。因为回归结果很少会像人们所期待的，因此，经常需要发展其他模型。例如，评价结果可能暗示方程遗漏了某个重要变量。在这种情况下，应该返回到第1步去查阅文献，在方程中增加合适的变量。然后重新按顺序完成每个步骤，直到第5步完成新方程的估计。只有对估计出的方程感到满意后，才能进入第6步。不过，不要太快做出这种判断，因为并不能仅仅为了拟合数据而去调整理论。研究者在改进方程时，必须小心谨慎，寻求合理的改进正是应用计量经济学的魅力所在。

最后，为了检验结果是否稳定，有必要估计其他设定形式的方程，这种方法称为敏感性分析，6.4节将对其进行详细讨论。

3.1.6 步骤6：报告结果

下面是报告回归分析估计结果的标准格式：

$$\begin{aligned} \hat{Y}_i &= 103.40 + 6.38X_i \\ &\quad (0.88) \\ &\quad t = 7.22 \\ N &= 20 \quad \bar{R}^2 = 0.73 \end{aligned} \tag{3-3}$$

括号中的数值表示估计参数的标准差，t值用来检验参数的真值不为零这个假设。这些衡量回归质量的各种方法将在接下来的章节中陆续讨论。[31] 值得注意的是，利用简单易懂的方式来报告回归结果是回归分析的重要组成部分。对于时间序列数据集，还应报告数据的频数（如季度或年度）以及样本区间。

大多数计算机程序可以把统计数值计算到小数点后8位或更高，但区分小数点后的位数和有效数字是很重要的，一般要求有效数字至少是2位或3位。

输出结果最重要的部分是对回归模型、假设条件、回归过程以及所用数据的说明。计量经济学报告必须包含足够多的信息以使他人可以完全理解整个研究。[32]较短的定义应该随方程给出，除非所使用的变量事先在表格中定义了。如果用到的是一组回归方程，表格中必须给出各个方程的相关信息。对于所有的数据操作以及数据来源必须进行详尽说明。当需要说明和解释的内容较多时，通常把这些内容整理为数据附件。在一般情况下，如果无法获得回归分析中所使用的数据，或只有在计算后才能得到，这个数据集本身也应列入附件中。

---

31 系数的标准差及t值将分别在4.2节及5.2节中详细说明。

32 例如，*Journal of Money, Credit, and Banking* 就要求作者提交其可以验证回归结果的真实数据集。请参见W.G. Dewald et al,“Replication in Empirical Economics,”*American Economic Review*, Vol. 76, No. 4, pp. 587-603，以及Daniel S. Hamermesh,“Replication in Economics,”*NBER Working paper 13026*, April 2007。

3.2 回归分析实例：餐厅选址

为了强化读者对应用回归分析六个基本步骤的理解，我们将展示一个完整的回归分析实例。为了确定Woody's餐厅[33]（Woody's是一个价格适中、24小时营业的家庭式连锁餐厅）下一个连锁店的最佳位置，有研究者决定建立回归模型来描述各个连锁店的总销售量。每家连锁店的总销售量都是地理位置相关属性的函数，如果可以找到描述这种函数关系的合理方程，那么，就可以用这个方程去帮助Woody's餐厅决定在什么地方修建新的连锁店。只要给出有关土地成本、建设成本及当地建筑和餐厅的市政法规数据，Woody's餐厅的老板就可以做出一个明智的决定。

（1）查阅文献，建立理论模型。阅读有关餐饮业的文献，但最主要还是和公司里的专家们交谈。他们会给出Woody's餐厅理想地址应具备的属性。专家们会说，所有连锁店的特征都是一样的（的确，这其实就是对连锁店的批判），都身处“郊区，零售或住宅区”环境中（即既不在市中心，也不在乡村）。正因为如此，可以认为许多影响其他连锁店销量的因素并不适用于本例，因为所有Woody's餐厅的位置都是相似的。（如果把Woody's餐厅和其他连锁店做比较，那么这些变量就是合适的。）

另外，Woody's餐厅战略规划部的人提出的观点也值得重视，他们认为地理位置间的价格差异与消费差别不及某个位置的独特性重要。这点引起了研究者的重视，因为最初考虑的变量（总销售额）会因各个地方价格的变化而变化。由于公司会控制价格，而需要估计的是“潜在”销售量，所以，已经存在的连锁店的顾客数量作为被解释变量。研究数据来自最近几年Woody's餐厅服务员开出的发票或账单。

（2）设定模型：选择解释变量及函数形式。经过上面的准备，可能会归纳出若干可供选择的解释变量。仔细分析后发现，实际上只有三个主要因素决定销售量。分别是：在连锁店附近居住的人口密度、当地居民的一般收入水平以及附近的直接竞争对手的数量。另外，还有两个较好的潜在解释变量：一是每天经过当地的车辆数，二是连锁店的营业时间。经过认真思考，决定舍弃最后两个变量，其理由是各个地方的连锁店都已经有足够长的营业时间，从而有稳定的顾客量，另外，搜集经过各个地方的车辆数的成本非常高。如果当地居住的人口数不能很好地衡量潜在消费者的数量，那就申请经费，用于搜集详尽的路过车辆的数据。

最终确定的解释变量：

N——竞争，当地Woody's连锁店的方圆2英里内的直接竞争对手数量
P——人口，当地Woody's连锁店的方圆3英里内的居住人口数
I——收入，变量P中度量的居住人口的平均收入水平

因为没有理由怀疑线性函数形式和古典随机误差项，于是决定直接选用它们。

（3）假设参数的预期符号。当决定应该包括哪些变量后，假设参数的符号就变得很容易了。其中两个变量的符号很容易确定。一是竞争越大，顾客就越少（该地区的人口和收入水平都给定不变的前提下）；二是居住在这家连锁店附近的人口越密集，顾客就会越多（竞争和收入水平保持不变）。研究者或许会认为某个地方的收入水平越高，在家庭式餐厅就餐的人就会越多。然而，对高收入地区的人而言，高收入群体可能希望到更高级的餐厅而非到Woody's一样的家庭式餐厅就餐。因此，收入变量对Woody's餐厅的销售量可能只有很小的促进作用。综合以上讨论，预期符号为：

$$Y_i = f(\overset{-}{N_i}, \overset{+}{P_i}, \overset{+?}{I_i}) + \varepsilon_i = \beta_0 + \beta_N N_i + \beta_P P_i + \beta_I I_i + \varepsilon_i \tag{3-4}$$

式中，每个变量上面的符号表示在其他两个解释变量保持不变的情况下，该解释变量对被解释

33 该案例中的数据都是真实的（它们来自于在南加利福尼亚州33家Denny's餐厅的样本），但是解释变量个数比实际研究中少很多，数据文件名为WOODY3。

变量的预期影响，ε_i是满足古典假设的随机误差项。

（4）搜集、检查并整理数据。本研究覆盖了Woody's餐厅的每家连锁店，得到了33个位置的被解释变量与解释变量。对数据进行检查后，有三个原因对数据的质量抱有信心：同一个变量在不同的餐厅采用相同的口径测量，样本中包含了所有连锁店，所有的数据都来自同一年度。在计算机运行EViews软件所使用的样本数据以及获得的回归估计结果如表3-1、表3-2所示，运行Stata软件所使用的样本数据以及获得的回归估计结果如表3-3、表3-4所示。

表3-1 Woody's 餐厅案例数据（应用EViews 软件）

obs	Y	N	P	I
1	107919.0	3.000000	65044.00	13240.00
2	118866.0	5.000000	101376.0	22554.00
3	98579.00	7.000000	124989.0	16916.00
4	122015.0	2.000000	55249.00	20967.00
5	152827.0	3.000000	73775.00	19576.00
6	91259.00	5.000000	48484.00	15039.00
7	123550.0	8.000000	138809.0	21857.00
8	160931.0	2.000000	50244.00	26435.00
9	98496.00	6.000000	104300.0	24024.00
10	108052.0	2.000000	37852.00	14987.00
11	144788.0	3.000000	66921.00	30902.00
12	164571.0	4.000000	166332.0	31573.00
13	105564.0	3.000000	61951.00	19001.00
14	102568.0	5.000000	100441.0	20058.00
15	103342.0	2.000000	39462.00	16194.00
16	127030.0	5.000000	139900.0	21384.00
17	166755.0	6.000000	171740.0	18800.00
18	125343.0	6.000000	149894.0	15289.00
19	121886.0	3.000000	57386.00	16702.00
20	134594.0	6.000000	185105.0	19093.00
21	152937.0	3.000000	114520.0	26502.00
22	109622.0	3.000000	52933.00	18760.00
23	149884.0	5.000000	203500.0	33242.00
24	98388.00	4.000000	39334.00	14988.00
25	140791.0	3.000000	95120.00	18505.00
26	101260.0	3.000000	49200.00	16839.00
27	139517.0	4.000000	113566.0	28915.00
28	115236.0	9.000000	194125.0	19033.00
29	136749.0	7.000000	233844.0	19200.00
30	105067.0	7.000000	83416.00	22833.00
31	136872.0	6.000000	183953.0	14409.00
32	117146.0	3.000000	60457.00	20307.00
33	163538.0	2.000000	65065.00	20111.00

相关系数矩阵

	Y	N	P	I
Y	1.000000	-0.144225	0.392568	0.537022
N	-0.144225	1.000000	0.726251	-0.031534
P	0.392568	0.726251	1.000000	0.245198
I	0.537022	-0.031534	0.245198	1.000000

表3-2 计算机实际输出结果（应用EViews 软件）

Dependent Variable: Y
Method: Least Squares
Date: 02/29/09 Time: 14:55
Sample: 1 33
Included observations: 33

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	102192.4	12799.83	7.983891	0.0000
N	-9074.674	2052.674	-4.420904	0.0001
P	0.354668	0.072681	4.879810	0.0000
I	1.287923	0.543294	2.370584	0.0246

R-squared	0.618154	Mean dependent var	125634.6
Adjusted R-squared	0.578653	S.D. dependent var	22404.09
S.E. of regression	14542.78	Akaike info criterion	22.12079
Sum squared resid	6.13E+09	Schwarz criterion	22.30218
Log likelihood	-360.9930	F-statistic	15.64894
Durbin-Watson stat	1.758193	Prob(F-statistic)	0.000003

obs	Actual	Fitted	Residual	Residual Plot
1	107919.	115090.	-7170.56	
2	118866.	121822.	-2955.74	
3	98579.0	104786.	-6206.86	
4	122015.	130642.	-8627.04	
5	152827.	126346.	26480.5	
6	91259.0	93383.9	-2124.88	
7	123550.	106976.	16573.7	
8	160931.	135909.	25021.7	
9	98496.0	115677.	-17181.4	
10	108052.	116770.	-8718.09	
11	144788.	138503.	6285.43	
12	164571.	165550.	-979.034	
13	105564.	121412.	-15848.3	
14	102568.	118275.	-15707.5	
15	103342.	118896.	-15553.6	
16	127030.	133978.	-6948.11	
17	166755.	132868.	33886.9	
18	125343.	120598.	4744.90	
19	121886.	116832.	5053.70	
20	134594.	137986.	-3391.59	
21	152937.	149718.	3219.43	
22	109622.	117904.	-8281.51	
23	149884.	171807.	-21923.2	
24	98388.0	99147.7	-759.651	
25	140791.	132537.	8253.52	
26	101260.	114105.	-12845.4	
27	139517.	143412.	-3895.30	
28	115236.	113883.	1352.60	
29	136749.	146335.	-9585.91	
30	105067.	97661.9	7405.12	
31	136872.	131544.	5327.62	
32	117146.	122564.	-5418.45	
33	163538.	133021.	30517.0	

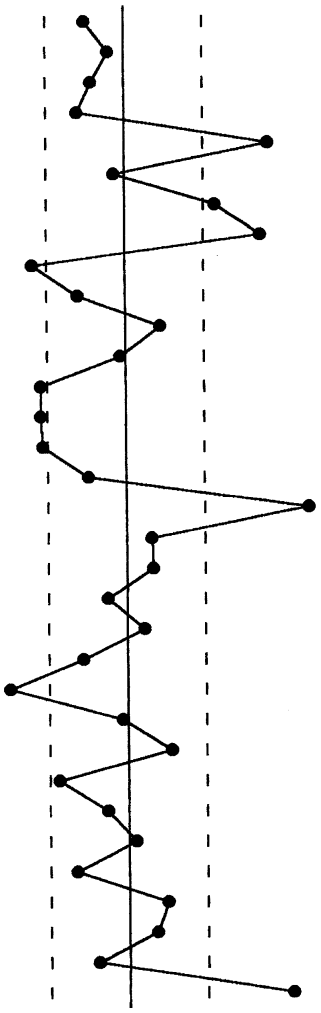

表3-3 Woody's 餐厅案例数据（应用Stata软件）

	Y	N	P	I		Y	N	P	I
1.	107919	3	65044	13240	18.	125343	6	149894	15289
2.	118866	5	101376	22554	19.	121886	3	57386	16702
3.	98579	7	124989	16916	20.	134594	6	185105	19093
4.	122015	2	55249	20967	21.	152937	3	114520	26502
5.	152827	3	73775	19576	22.	109622	3	52933	18760
6.	91259	5	48484	15039	23.	149884	5	203500	33242
7.	123550	8	138809	21857	24.	98388	4	39334	14988
8.	160931	2	50244	26435	25.	140791	3	95120	18505
9.	98496	6	104300	24024	26.	101260	3	49200	16839
10.	108052	2	37852	14987	27.	139517	4	113566	28915
11.	144788	3	66921	30902	28.	115236	9	194125	19033
12.	164571	4	166332	31573	29.	136749	7	233844	19200
13.	105564	3	61951	19001	30.	105067	7	83416	22833
14.	102568	5	100441	20058	31.	136872	6	183953	14409
15.	103342	2	39462	16194	32.	117146	3	60457	20307
16.	127030	5	139900	21384	33.	163538	2	65065	20111
17.	166755	6	171740	18800					

(Obs=33)	Y	N	P	I
Y	1.0000			
N	-0.1442	1.0000		
P	0.3926	0.7263	1.0000	
I	0.5370	-0.0315	0.2452	1.0000

表3-4 计算机实际输出结果（应用Stata 软件）

Source	SS	df	MS
Model	9,9289e+09	3	3.3096e+09
Residual	6.1333e+09	29	211492485
Total	1.6062e+10	32	501943246

Number of obs=**33**
F(3, 29)=**15.65**
Prob>F=**0.0000**
R-squared=**0.6182**
Adj R-squared=**0.5787**
Root MSE=**14543**

Y	Coef.	Std.Err.	t	P>|t|	[95% Conf.	Interval]
N	-9074.674	2052.674	-4.42	0.000	-13272.86	-4876.485
P	.3546684	.0726808	4.88	0.000	.2060195	.5033172
I	1.287923	.5432938	2.37	0.025	.1767628	2.399084
_cons	102192.4	12799.83	7.98	0.000	76013.84	128371

	Y	Yhat	residu-s		Y	Yhat	residu-s
1.	107919	115089.6	115089.6	8.	160931	135909.3	135909.3
2.	118866	121821.7	121821.7	9.	98496	115677.4	115677.4
3.	98579	104685.9	104685.9	10.	108052	116770.1	116770.1
4.	122015	130642	130642	11.	144788	138502.6	138502.6
5.	152827	126346.5	126346.5	12.	164571	165550	165550
6.	91259	93383.88	93383.88	13.	105564	121412.3	121412.3
7.	123550	106976.3	106976.3	14.	102568	118275.5	118275.5

（续）

	Y	Yhat	residu-s		Y	Yhat	residu-s
15.	103342	118895.6	118895.6	25.	140791	132537.5	132537.5
16.	127030	133978.1	133978.1	26.	101260	114105.4	114105.4
17.	166755	132868.1	132868.1	27.	139517	143412.3	143412.3
18.	125343	120598.1	120598.1	28.	115236	113883.4	113883.4
19.	121886	116832.3	116832.3	29.	136749	146334.9	146334.9
20.	134594	137985.6	137985.6	30.	105067	97661.88	97661.88
21.	152937	149717.6	149717.6	31.	136872	131544.4	131544.4
22.	109622	117903.5	117903.5	32.	117146	122564.5	122564.5
23.	149884	171807.2	171807.2	33.	163538	133021	133021
24.	98388	99147.65	99147.65				

（5）方程的估计与评价。获得数据集并录入计算机后，就可用普通最小二乘法进行回归分析。但开始之前，必须再次检查模型是否存在理论错误。直到自己认为没有问题为止，即便可能尚有一些不确定。于是，估计方程，得到如下结果：

$$\begin{aligned} \hat{Y}_i &= 102\,192 - 9\,075N_i + 0.355P_i + 1.288I_i \\ &\quad (2\,053) \quad (0.073) \quad (0.543) \\ t &= -4.42 \quad 4.88 \quad 2.37 \\ N &= 33 \quad \bar{R}^2 = 0.579 \end{aligned} \tag{3-5}$$

从短期来看，这个方程满足要求。尤其是方程中参数估计值的符号与预期相同。虽然模型的整体拟合优度不是很好，但考虑到这些连锁店所在位置差异较大，这个结果应该说是合理的。为了预测Y值，将得到的每个连锁店预设位置的N、P及I值代入方程（3-5）。如果排除其他因素的影响，则对于Woody's餐厅来说，Y的预期值越高，位置就越好。

（6）报告结果。方程（3-5）归纳的结果已经可以满足报告的要求。（注意：虽然在第5章之前不会涉及估计参数的标准差和t统计量[34]，但还是包含了它们的完整信息。）然而，对初学者来说，从回归分析输出结果中找到所有需要报告的数据并非易事。如果能花些时间仔细阅读表3-1～表3-4中Woody's餐厅模型的计算机输出表格和信息，那么，在读取自己计算机的输出结果时，就可能会比较轻松。这里输出的表格是由EViews和Stata软件在计算机中生成的，那些由SAS、SHAZAM、TSP及其他软件生成的表格也与之类似。

第一项列出的是实际数据，随后是数据集中所有变量之间的简单相关系数，接下来就是一系列的估计参数，它们的估计标准差以及t统计量。紧接着是判定系数R^2、调整的判定系数$\bar{R}^2$、回归标准差、残差平方和RSS、F值以及将在后面章节中介绍的其他信息。最后，列出Y_s的观察值和预测值以及残差，绘出了残差的图形。数值后面紧跟的“E+06”或“E-01”表示数值采用科学标记法，意味着数值显示的小数点必须向右移动6位或向左移动1位。

在以后的章节中，还会回到这个例子，将新学到的各种检验方法和思想应用进去。

小结

1. 通常，确定被解释变量后，应用回归分析采用6个步骤：

34 在本书中，估计参数下方的括号中的数值是估计参数的标准差。由于有些作者会把t值放入括号中，所以在阅读期刊文章或其他书籍时要特别注意。

a. 查阅文献并建立理论模型。
b. 确定模型，选择解释变量及函数形式。
c. 假设参数的预期符号。
d. 搜集、检查和整理数据。
e. 估计与评价方程。
f. 报告结果。

2. 虚拟变量只能取值0或1，这取决于是否满足某些特定状态条件。虚拟变量的例子就是当为女性时X等于1，为男性时X等于0。

习题

（习题2的答案见附录A。）

1. 不查阅书本（或笔记），给出下列术语的定义，然后与书本上的相比较。
   a. 应用回归分析中的6个步骤
   b. 虚拟变量
   c. 截面数据
   d. 设定误差
   e. 自由度
2. 虚拟变量不像它的名字听起来好理解，在没有一定练习的情况下并不容易掌握。
   a. 设定一个虚拟变量用以区别在计量经济学班级中的本科生与研究生。
   b. 设定一个回归方程来解释班上每个同学第1次计量经济学测试成绩（Y，满分4.0）作为学生以前学习统计学的成绩（G）、课堂学习时间（H）以及上面设定的虚拟变量（D）的函数。方程中还需要再加入其他变量吗？请说明理由。
   c. 虚拟变量D的参数符号的假设是什么？这个符号有赖于定义D的确切方法吗？（提示：特别地，假设在a部分的答案中你弄反了1和0的定义），弄反了会怎么样？
   d. 假设通过搜集的数据进行了回归分析，得到了D的参数估计值，这个参数的符号与你预期的符号一致，参数估计值的绝对值为0.5。在现实生活中这意味着什么？另外，如果班级中只有本科生或只有研究生又会怎样？
3. 是否文科学院中经济学家的薪水要比其他教授多？为了调查这个问题，观测由2 929个教员组成的规模比较小的学院中的样本，构造了并估计由4个变量解释薪水的模型：

$$\hat{S}_i = 36\,721 + 817M_i + 426A_i + 406R_i + 3\,559T_i + \cdots \quad (3\text{-}6)$$
$$(259)\quad (456)\quad (24)\quad (458)$$
$$\bar{R}^2 = 0.77 \qquad N = 2\,929$$

式中，S_i代表学院中第i位教授的薪水；M_i代表虚拟变量，如果第i位教授为男性时为1，否则为0；A_i代表虚拟变量，如果第i位教授为非裔美国人则为1，否则为0；R_i代表第i位教授的级别；T_i代表虚拟变量，如果第i位教授若是讲经济学类课程的教授则为1，否则为0。
   a. 请详细说明M的参数的含义。
   b. 该方程意味着在其他条件不变的情况下，非裔美国人比其他民族成员多赚426美元。这个参数的符号与你预期的一致吗？为什么？
   c. R是虚拟变量吗？如果不是，那它是什么？请详细说明R参数的含义。（提示：通常，教授随级别上升，薪水也随之增加。）
   d. 你的结论是什么？文科学院中经济学家的薪水要比其他教授更高吗？请说明理由。
   f. 事实上，方程以符号“+...”结尾，意味着不止4个解释变量。如果你在方程中要增加一个变量，会增加什么？请说明理由。
4. 回到Woody's餐厅案例的回归结果。
   a. 在任何应用回归中，极有可能忽略掉某个重要的解释变量。重新阅读本章中关于如何选择解释变量的论述，并在原有的模型中加入一个合理的解释变量（已经提到的变量除外）。为什么原模型中没

有加入这个变量？

b. 对于模型所选择的样本或解释变量，你有什么异议吗？

5. 假设在方程（3-5）中有关交通的数据仍然很难获得。但是有关交通的变量T_i还是可以定义成：如果快餐店前面的交通拥挤则为1，否则为0。进一步假设当新变量(T_i)被加入到方程中，得到回归结果为

$$\hat{Y}_i = 95\,236 - 7\,307N_i + 0.320P_i + 1.288I_i + 10\,994T_i$$
$$(2\,153) \quad (0.073) \quad (0.51) \quad (5\,577)$$
$$t = -3.39 \quad 4.24 \quad 2.47 \quad 1.97$$
$$N = 33 \qquad \bar{R}^2 = 0.617 \qquad (3\text{-}7)$$

a. 新变量参数的预期符号是什么？

b. 相较于原方程，你更喜欢这个方程吗？为什么？

c. 方程（3-7）中调整的判定系数$\bar{R}^2$更高是否意味着它比方程（3-5）更好？

6. 假设在3.2节中的有关人口的变量被定义为：P_i代表人口，当地 Woody's连锁店方圆3.5英里内的居住人口数（单位：1 000人）。

a. 假定P被这么定义，那么方程（3-5）中估计参数的斜率会变成什么？

b. 假定P被这么定义，那么在方程（3-7）中估计参数的斜率会变成什么？

c. 这种变化会影响到常数项的估计值吗？

7. 利用EViews、Stata或者你自己拥有的软件，运用表3-1的数据，估计方程（3-5），你能得到同样的结果吗？

8. 研究生入学考试（GRE）中，经济学科目考试是一种经济学知识和分析能力的综合测试，它作为学生申请“沉闷科学”的博士入学条件。这些年来，有人指出GRE像学生能力倾向测试（SAT）一样，对女性和某些种族存在偏见。为了检验GRE经济学科目考试是否对女性存在偏见，Mary Hirschfeld、Robert Moore以及Eleanor Brown 估计了下面的方程（括号中的数值为标准差）[35]

$$\widehat{GRE}_i = 172.4 + 39.7G_i + 78.9GPA_i + 0.203SATM_i + 0.110SATV_i$$
$$(10.9) \quad (10.4) \quad (0.071) \quad (0.058)$$
$$N = 149 \qquad \bar{R}^2 = 0.46 \qquad (3\text{-}8)$$

式中，GRE_i代表第i个学生在GRE经济学科目中的考试分数；G_i代表虚拟变量，如果第i个学生是男性则为1，否则为0；GPA_i代表第i个学生经济学课程的GPA（A=4, B=3，等等）；$SATM_i$代表第i个学生在学生能力倾向测试中的数学成绩；$SATV_i$代表第i个学生在学生能力倾向测试中的其他部分的成绩。

a. 详细说明在这个方程中，G_i参数的含义（提示：详细说明39.7代表什么？）

b. 这个结果能够证明GRE对女性存在偏见吗？为什么？

c. 如果在方程（3-8）中增加一个变量，你会选择增加什么变量？请说明理由。

d. 假设作者已经定义他们的性别变量就是G_i，为虚拟变量，如果第i个学生是女性则为1，否则为0。那么在这种情形下，方程（3-8）会怎么变化？（提示：只有截距和这个虚拟变量的参数发生变化。）

9. Michael Lovell [36]估计了不同型号汽车的汽油里程模型（括号中的数值为标准差）。

$$\hat{G}_i = 22.008 - 0.002W_i - 2.76A_i + 3.28D_i + 0.415E_i$$
$$(0.001) \quad (0.71) \quad (1.41) \quad (0.097)$$
$$\bar{R}^2 = 0.82$$

式中，G_i代表消费者联盟公布的在真实道路上测试的第i种车型每加仑汽油所能行驶的距离（单位：英里）W_i代表第i种车型的总重量（单位：英磅）；A_i代表虚拟变量，如果第i种车型是自动变速器则为1，否则为0；D_i代表虚拟变量，如果第i种车型采用柴油发动机则为1，否则为0；E_i代表美国环境保护局估计的第i种车型每加仑汽油所能行驶的距离（单位：英里）。

a. 假设W和E的参数符号。估计结果中，如果存在参数符号与预期不一致，那么会

35 Mary Hirschfeld , Robert L. Moore, and Eleanor Brown, “Exploring the Gender Gap on the GRE Subject Test in Economics,” *Journal of Economic Education*, Vol. 26, No.1, pp. 3-15.

36 Michael C.Lovell, “Test of the Rational Expectations Hypothesis,” *American Economic Review*, Vol. 76, No. 1, pp. 110-124.

是哪些变量的参数？

b. 详细说明A_i和D_i回归参数的含义。（提示：记住E在方程中。）

c. Lovell 的模型中包含了用于检验特定假设的变量，但这个变量在其他研究汽油里程的模型中并不需要。你认为Lovell增加的变量是哪一个？Lovell想检验的特定假设是什么？

10. 老板打电话告诉你，他打算开拍他最新的破票房之作——《经济学家入侵》（第二部）（*Invation of the Economists, Part* Ⅱ），并要求你建立一个过去五年里所有电影总收入的模型。你的模型是（括号中的数值为标准差）[37]

$$\hat{G}_i = 781 + 15.4T_i - 992F_i + 1\,770J_i + 3\,027S_i - 3\,160B_i + \cdots$$
$$(5.9) \quad (674) \quad (800) \quad (1\,006) \quad (2\,381)$$
$$\bar{R}^2 = 0.485 \quad N = 254$$

式中，G_i代表第i部电影的最后总收入（单位：1 000美元）；T_i代表第i部电影在上映的第一周，播放它的影院数；F_i代表虚拟变量，如果第i部电影的明星是女性则为1，否则为0；J_i代表虚拟变量，如果第i部电影在6月或7月放映则为1，否则为0；S_i代表虚拟变量，如果第i部电影的明星是超级巨星（比如Tom Cruise 或者 Milton）则为1，否则为0；B_i代表虚拟变量，如果第i部电影至少有一个配角是超级巨星则为1，否则为0。

a. 假设每个斜率参数的符号。估计结果中，如果存在参数符号与预期不同，那么，会是哪些变量的参数？

b. 出演《经济学家入侵》第一部的演员Milton向老板要价400万美元拍摄续集。如果估计是可信的，那么老板应该答应Milton，还是用50万美元雇用一个无人知晓的演员Fred呢？

c. 老板想要保持低成本，增加200次放映将会花费120万美元。假设你的模型可信，那么，老板应该增加放映次数吗？

d. 这部电影被安排在9月放映，在保证电影质量的情况下，为了能在7月上映，将会花费100万美元来加速拍摄。仍然假设你的模型可信，那么加速拍摄值得吗？

e. 你一直在假设估计是可信的，但是否存在证据表明事实并非如此？解释你的答案。（提示：假设方程并不存在设定误差。）

11. 让我们来多做一些关于应用回归分析中的6个步骤的练习。假如想在eBay上购买一款苹果iPod（不论新旧），但又不希望出价过高。基于以前iPod的拍卖价格，建立回归模型是一种能够深入了解iPod竞标价格的方法[38]。

第1步是查阅文献。很幸运，你找到了一些好素材，尤其是Leonardo Rezende [39]在2008年撰写的一篇分析eBay、估计iPod价格模型的文章。

第2步是确定方程的解释变量及函数形式。由于有些是新iPod，有些是二手货但无瑕疵，还有些则是二手货且有刻痕或其他缺陷，因此，你遇到一个问题，即想在方程中包含一个度量iPod使用情况的变量。

a. 设定一个（或多个）变量，可以量化iPod三种不同的使用情况。在继续下一步之前请先回答这个问题。

b. 第3步是假设方程中参数的符号。如果方程设定如下，预期变量*NEW*、*SCRATCH*和*BIDRS*的符号各是什么？请说明理由。

$$PRICE_i = \beta_0 + \beta_1 NEW_i + \beta_2 SCRATCH_i + \beta_3 BIDRS_i + \varepsilon_i$$

式中，$PRICE_i$代表第i个iPod在ebay上的销售价格；NEW_i代表虚拟变量，如果第i个iPod是新的则为1，否则为0；$SCRATCH_i$代表虚拟变量，如果第i个iPod表面上有微小的瑕疵则为1，否则为0；$BIDRS_i$代表对第i个iPod报价的人数。

---

37 这个估计方程（并非这个问题）是来自哈佛商学院的一次管理经济学期末考试。

38 这是另一个hedonic模型的例子，其中，价格为被解释变量，解释变量是被解释变量的影响因素。如果想进一步了解hedonic模型，请参见11.8节。

39 Leonardo Rezende, "Econometrics of Auctions by Least Squares," *Journal of Applied Econometrics*, November / December 2008, pp. 925-948.

c. 第4步是搜集数据。幸运的是，Rezende拥有215个银色、4GB可上网的迷你苹果iPod的数据，所以他迫切地希望下载数据，进行初次回归分析。然而在进行之前，有人指出iPod拍卖在三星期前已经开始，他担心由于数据来自不同时间段而造成数据没有可比性。这种担忧有意义吗？为什么？

d. 第5步是利用Rezende的数据估计方程，得到：

$$\widehat{PRICE}_i = 109.24 + 54.99NEW_i - 20.44SCRATCH_i + 0.73BIDRS_i$$

	(5.34)	(5.11)	(0.59)
t =	10.28	-4.00	1.23

$N = 215$

参数估计值的符号与你的预期一致吗？请说明理由。

e. 第6步是报告结果。查看d部分中的回归分析结果，是否遗漏了一些应该报告的项目？

f. （选做）自己估计方程（数据文件名为IPOD3），说明在e部分的答案中你认为遗漏的项目的作用。