卓越大学出版联盟
学术出版精品工程

实用统计技术丛书

线性回归模型应用及判别

LINEAR REGRESSION MODELS AND DIAGNOSTICS

李元章　何春雄　著

华南理工大学出版社
SOUTH CHINA UNIVERSITY OF TECHNOLOGY PRESS

实用统计技术丛书

线性回归模型应用及判别

李元章　何春雄　著

华南理工大学出版社
SOUTH CHINA UNIVERSITY OF TECHNOLOGY PRESS
·广州·

内 容 提 要

本书介绍利用SAS软件进行回归分析的各种主题，讨论和演示如何处理数据，如何选择适当的模型，如何诊断模型，如何解释计算结果等. 适于了解概率论与数理统计基础知识和具有使用计算机软件基本经验的读者阅读. 可作为概率论与数理统计专业的研究生或数学专业高年级本科生的选修课参考教材，也适于用数理统计方法从事社会科学、自然科学研究的人员参考.

图书在版编目(CIP)数据

线性回归模型应用及判别/李元章，何春雄著. —广州：华南理工大学出版社，2016.5
(实用统计技术丛书)
ISBN 978-7-5623-4875-7

Ⅰ.①线… Ⅱ.①李… ②何… Ⅲ.①线性回归-回归分析-应用软件 Ⅳ.①O212.1

中国版本图书馆CIP数据核字(2016)第025722号

线性回归模型应用及判别
李元章 何春雄 著

出 版 人：卢家明
出版发行：华南理工大学出版社
(广州五山华南理工大学17号楼，邮编510640)
http://www.scutpress.com.cn E-mail:scutc13@scut.edu.cn
营销部电话：020-87113487 87111048(传真)
策划编辑：詹志青
责任编辑：詹志青
印 刷 者：佛山市浩文彩色印刷有限公司
开 本：787mm×1092mm 1/16 印张：19.5 字数：487千
版 次：2016年5月第1版 2016年5月第1次印刷
印 数：1～1000册
定 价：41.00元

前　言

回归分析是推断统计的重要分支之一，其主要的研究内容是分析和刻画随机变量之间的关联性．本书的内容包括应用 SAS 软件进行回归分析的各种主题，探讨和演示如何处理数据，如何选择适当的模型，如何诊断模型，如何解释计算结果．本书的重点是“数据分析”，而不是回归分析理论，因此要求读者至少学过一门包括回归分析在内的数理统计课程，并有一本回归分析的教材做参考，进而运用回归分析知识，利用 SAS 软件工具来实现、理解和解释回归分析．

本书的内容分为 9 章．第 1 章简要叙述线性回归的有关知识，包括参数估计、模型诊断(假设检验)和预测等．第 2 章介绍 SAS 软件工具的基本功能和使用的入门知识．第 3 章则通过一个例子，比较系统地介绍如何利用 SAS 软件工具实现回归分析，包括数据检查的各种手段、直观分析、变元变换和结果分析等．第 4 章专门针对分类数据，介绍回归分析的建模方法和计算结果的解释方法．第 5 章介绍回归分析模型诊断的各种实用技术，包括异常数据诊断、共线性诊断、正态性检验、独立性检验等．第 6 章讨论广义线性回归模型的参数估计、模型诊断等问题，该模型中的预测变量可以是分类型的，也可以是连续型的．第 7 章介绍多元方差分析和协方差分析模型，讨论该类模型的基本假设和用于对广义线性模型做检验的具体方法．第 8 章介绍用 SAS 分析重复测量数据的方法．具体介绍使用 SAS 的程序 GLM 分析重复测量数据的常规方法，阐述这种常规方法的基本假设和局限性，并且探讨使用 SAS 的程序 mixed(混合模型)重复测量数据的分析方法．第 9 章讨论预测变量的共线性问题，包括共线性对建模的影响、共线性识别直观分析和统计检验以及共线性问题的处理方法等．需要特别说明的是，由于 SAS 的内部函数名不区分字母大小写，因此本书 SAS 程序中会出现字母大小写混用的现象，而在行文中 SAS 的内部函数名、变量名都用小写，变量名一般用小写，英文缩略词用大写．

李元章教授曾在美国某研究院、兰州大学、华南理工大学等多地讲授过本书的内容，在实际问题的研究中用到本书中涉及的多种模型，并作为华南理工大学的客座教授，与何春雄教授合作，在华南理工大学举办实用统计技术系列讲座．本书内容是在讲座“回归分析与 SAS 应用”基础上加工整理而成的．在此，我们特别感谢华南理工大学数学学院和研究生院的大力支持，并衷心感谢华南理工大学出版社的精诚合作．由于作者水平有限，难免会有疏漏或不当之处，恳请同行和读者指正．

编　者

2016 年 1 月

目　录

1 回归分析

做检验本书的内容包括应用 SAS 软件进行回归分析的各种主题. 需要强调的是，本书是关于“数据分析”的，探讨和演示如何处理数据、如何选择适当的模型、如何诊断模型、如何解释计算结果. 我们假定读者至少学过一门包括“回归分析”在内的数理统计课程，并有一本《回归分析》的书做参考，本书旨在运用回归分析知识，结合 SAS 软件工具，去实现、理解和解释回归分析.

众所周知，统计学是关于数据收集、分析、解释和表述的科学. 统计学在多种科学领域有着广泛的应用，从自然科学到社会科学，从商业到政府、医疗、工业部门等，几乎所有的行业都需要统计工具，统计学的技能使我们能够合理地收集、分析和解释数据，并由此作相应的决策. 统计学的概念使人们能够在不同程度和范围内解决问题；统计学的思想使人们能在掌握事物的本质的前提下进行决策.

在通常情况下，统计分析有两类，一类是描述性的分析，它几乎应用于所有的领域. 描述性统计分析可以追溯到 17 世纪，那时统计学才刚刚诞生. 来自伦敦平民出身的约翰·格兰特(John Graunt)在翻阅由教区秘书主办的教堂周刊时，注意到其上列出的各教区出生的人数、接受洗礼仪式的人数和去世的人数. 这个被称为“死亡清单”的表上还列出了死亡的原因. 作为店主的格兰特用我们所说的描述性统计的方式整理了这些数据，并以《基于死亡清单的自然与政策观察报告》为题予以发表. 因此，格兰特很快被选为皇家科学协会成员. 正因为如此，统计学中借用了一些社会学的概念，比如“总体”(population)的概念. 由于起初统计学往往研究与人类行为有关的问题，因此有人认为统计学无法阐明物理学的精确性. 一般地说，描述性统计处理的是描述性问题：数据能否以有用的方式(数据的或图形的)进行总结或归纳，从而洞察人口中的某些问题？数据描述的最基本的例子是均值和标准差，图形式的总结包括各种图表和曲线图.

另一类比较复杂的统计分析是所谓的推断统计，它对数据的模式进行建模，考虑到随机性并从中推断大样本的性质. 这些推断可以是回答是或否的(如假设检验)、数字特征的估计(即参数估计)、未来观测值的预测、关联性的描述(相关性分析)以及对随机变量之间的关系建模(即回归分析)等等. 其它建模技术还有方差分析(analysis of variance，ANOVA)、时间序列分析和数据挖掘等等.

还需指出，分析因果关系的统计学研究方法还有两种类型，即试验研究和观测研究. 在这两种类型的研究中，独立变量(亦称预测变量或解释变量)对从属变量(亦称响应变量或应变量)的性态影响的差别可以观测到. 这两种研究类型的差别体现在具体的研究是如何进行的，每种都可以是很有效的. 试验研究关注被研究系统的量测值，操作该系统，并由新的量测值来确定量测值是否有所改善. 与此相反，观测研究不考虑实验的操作，而是收集数据，并研究预测变量与响应变量之间的相关性.

有多种重要的应用于不同领域的统计学方法，例如：

- t 检验：检验两个正态总体的均值是否相等；
- χ^2 检验：检验两个正态总体的方差是否相同；
- 方差分析：检验均值或效果是否相同；
- 曼-怀特(Mann-Whitney)U 检验：检验两个观测总体中位数是否相同；
- 回归分析：建立随机变量之间关系的模型，确定变量变化幅度的关系，并可基于所建模型进行预测；
- 相关性分析：刻画随机变量线性关系的强度和方向；
- 费舍(Fisher)的最小显著差异性检验：多重比较中均值差别的检验；
- 皮尔逊(Pearson)乘积－矩相关系数：度量来自同一对象的变量 X 与 Y 的线性关系的程度；
- 斯皮尔曼(Spearman)秩相关系数：两个变量之间相关程度的一个非参数度量.

本书主要介绍回归分析的知识. 本章将给出回归分析的基本知识的导言. 有关回归分析的详细介绍，请读者参考其它统计学教材.

1.1 一元线性回归及相关性

在统计学中，线性回归是一般回归分析的一种，其中一个或多个独立变化的量(亦称预测变量)与另一个变量(亦称响应变量或应变量)之间的关系用一个线性函数来刻画，我们称该函数为线性回归方程，其中用一个或多个参数对独立变量进行线性组合，这些参数称为回归系数. 只有一个独立变量的线性回归方程表示一条直线. 线性回归的结果需要进行统计分析. 关于线性回归的经典的假设包括样本(Y_i, X_i)以随机方式选自所研究的总体，应变量为实直线上的连续变量，误差项 ε 相互独立且同服从正态分布，也就是说误差为 i. i. d(即独立同分布)且为高斯(Gauss)分布. 值得注意的是，这些假设意味着误差项在统计意义下不依赖于预测变量 X_i，也就是说，ε 统计独立于预测变量 X_i，除非特别申明，本书都采用这些假设. 在较为现代的分析处理过程中，这些假设都可以放松，特别是关于误差项的正态分布假设往往不成立，除非样本容量较大，当样本容量较大时，由中心极限定理可知，只要误差项具有有限方差和不太强的相关性，即使误差项不服从正态分布，在参数估计时也可近似认为它服从正态分布.

对于一元线性回归，其假设和模型可以概括为下列(1)～(4).

(1)应变量(响应变量)都是数值型的(对于变量为分类变量的情形，将在第 4 章专门讨论).

(2)应变量之间是相互独立的.

(3)应变量的均值与解释变量之间关系近似为线性(直线)，且应变量在不同预测点的方差相同.

(4)模型为

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2). \tag{1.1}$$

在以上(1)～(4)的假设和模型下，一元线性回归模型可借助条件数学期望等价地写为

$$E[Y|X] = \beta_0 + \beta_1 X.$$

也就是说，在给定 X 的条件下，Y 的条件数学期望为 X 的线性变换．当然要注意，这一表达式在 X 已知时误差项的条件期望为0这一假设下才成立．实际上，本书以下的讨论中都假定预测变量为确定的，不是随机变量．

一元线性回归模型的直观意义可总结为如下的(1)～(4)．

(1) β_0，β_1 为未知参数．

(2) β_0 为 $X=0$ 时的应变量的均值(回归直线在 Y 轴上的截距)．

(3) β_1 为 X 增加一个单位时应变量的平均变化率(回归直线的斜率)．若 $\beta_1>0$，则 Y 与 X 正相关；若 $\beta_1<0$，则 Y 与 X 负相关；若 $\beta_1=0$，则 Y 与 X 不相关．

(4) $\beta_0+\beta_1 X$ 为解释变量取值 X 时应变量的均值．

设 $(x_i, y_i)(i=1, 2, \cdots, n)$ 为解释变量 X 与应变量 Y 的 n 对观测值，由此可绘出图1－1所示的散点图．

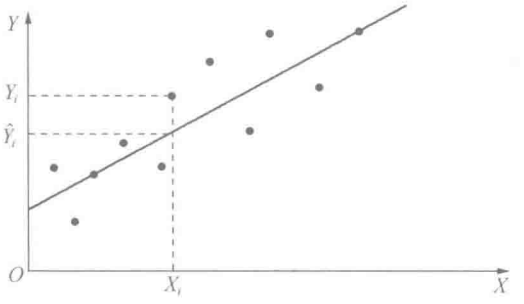

图1－1　解释变量 X 与应变量 Y 的散点图和回归直线示意图

我们以下的目标是由散点图1－1所示的解释变量 X 与应变量 Y 对应的观测值来选取(估计)参数 $\hat{\beta}_0$，$\hat{\beta}_1$，使得观测值与直线上点的均方误差(亦称残差平方和，SSE)达到最小，这就是所谓的最小二乘法．

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \quad \mathrm{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2. \tag{1.2}$$

如果残差平方和为0，则得到一条直线且所有点都在该直线上：如果残差平方和不为0，则所有点 (x_i, y_i) 不会落在同一条直线上，所以不得不移动直线，使其离某些点较近而离其它点较远．因此，拟合直线的过程为，先计算出各点的残差，进而算出残差平方和，再通过最小化残差平方和，最后选出最好的拟合直线(确定截距和斜率)．

可将最小二乘估计的过程想象成两个步骤，先画出应变量的均值点，再以该点为支点，旋转直线使残差达到最小．

1.1.1　模型的参数估计

对于预测变量的取值 $(x_1, x_2, \cdots, x_n)$ 及相应的应变量

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1,2,\cdots,n$$

的观测值 $(y_i, y_2, \cdots, y_n)$，记

$$Q(\beta_0,\beta_1) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2, \tag{1.3}$$

则对参数 β_0 和 β_1 的最小二乘估计就是求 $Q(\beta_0, \beta_1)$ 的最小值点 $\hat{\beta}_0$ 和 $\hat{\beta}_1$. 令式(1.3)右端对 β_0、β_1 的偏导数均为0，有

$$\begin{cases} n\beta_0 + (\sum_{i=1}^{n} x_i)\beta_1 = \sum_{i=1}^{n} y_i \\ (\sum_{i=1}^{n} x_i)\beta_0 + (\sum_{i=1}^{n} x_i^2)\beta_1 = \sum_{i=1}^{n} x_i y_i \end{cases}, \tag{1.4}$$

解未知数为 β_0 和 β_1 的二元一次方程(1.4)，得到

$$\begin{cases} \beta_1 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \dfrac{S_{xy}}{S_{xx}} \\ \beta_0 = \bar{y} - \beta_1 \bar{x} \end{cases}.$$

式中，$\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i, \bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$. 由于 $\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - \bar{x}) y_i$，可以得到 β_0 和 β_1 的最小二乘估计量为

$$\begin{cases} \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x}) Y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}. \\ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \end{cases} \tag{1.5}$$

为得到 σ^2 的估计，记 $S_e = \sum_{i=1}^{n} [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$，经简单计算知 $E[S_e] = (n-2)\sigma^2$，从而得到

$$\hat{\sigma}^2 = \frac{S_e}{n-2} \tag{1.6}$$

为 σ^2 的无偏估计.

由于假定误差项 $\varepsilon_i \sim N(0, \sigma^2), i = 1, 2, \cdots, n$ 且不相关，故 $Y_1, Y_2, \cdots, Y_n$ 相互独立且方差相等，从而可知式(1.5)给出的估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 也分别是 β_0 和 β_1 的最大似然估计. 另外，还可以证明，

$$S_e、\bar{Y}、\hat{\beta}_1 \text{ 相互独立，且} \frac{S_e}{\sigma^2} \sim \chi^2(n-2). \tag{1.7}$$

1.1.2 参数估计量的性质

为了下一步对模型的线性性作假设检验，我们需要探讨估计量 $\hat{\beta}_0$、$\hat{\beta}_1$ 及 $\hat{\sigma}^2$ 的分布. 注意到式(1.5)中 $(x_1, x_2, \cdots, x_n)$ 为确定数值，所以 $\hat{\beta}_1$ 是独立正态随机变量 $Y_1, Y_2, \cdots, Y_n$ 的线性组合，从而 $\hat{\beta}_1$ 服从正态分布，其均值为

$$E[\hat{\beta}_1] = E\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})Y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right] = \frac{\sum_{i=1}^{n}(x_i-\bar{x})E[Y_i]}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(\beta_0+\beta_1 x_i)}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \beta_1,$$

方差为

$$\mathrm{Var}[\hat{\beta}_1] = \mathrm{Var}\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})Y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right] = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2\mathrm{Var}[Y_i]}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2\sigma^2}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^2} = \frac{\sigma^2}{S_{xx}}.$$

总之，有

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right). \tag{1.8}$$

再来求 $\hat{\beta}_0$ 的分布. 首先,由于 $Y_1, Y_2, \cdots, Y_n$ 相互独立,容易推得 $\mathrm{Cov}(\bar{Y}, \hat{\beta}_1) = 0$,从而 $\bar{Y}$ 与 $\hat{\beta}_1$ 相互独立,所以 $\hat{\beta}_0$ 服从正态分布,其均值为

$$E[\hat{\beta}_0] = E[\bar{Y} - \hat{\beta}_1\bar{x}] = E[\bar{Y}] - \bar{x}E[\hat{\beta}_1] = \frac{1}{n}\sum_{i=1}^{n}E[Y_i] - \bar{x}\hat{\beta}_1$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\beta_0+\beta_1 x_i) - \bar{x}\beta_1 = \beta_0;$$

方差为

$$\mathrm{Var}[\hat{\beta}_0] = \mathrm{Var}[\bar{Y} - \hat{\beta}_1\bar{x}] = \mathrm{Var}[\bar{Y}] + \bar{x}^2\mathrm{Var}[\hat{\beta}_1]$$

$$= \frac{1}{n}\sigma^2 + \bar{x}^2\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2.$$

总之，又有

$$\hat{\beta}_0 \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2\right). \tag{1.9}$$

1.1.3 模型的统计推断

模型的统计推断包括参数的假设检验(双边或单边)和置信区间.

1. 关于 β_1 的统计推断

双边检验	单边检验
• H_0: $\beta_1 = 0$;	• H_0: $\beta_1 = 0$;
• H_A: $\beta_1 \neq 0$;	• H_A: $\beta_1 > 0$; (或 H_A: $\beta_1 < 0$;)
• 统计量: $t_{\mathrm{obs}} = \dfrac{\hat{\beta}_1}{\hat{\sigma}(\hat{\beta}_1)}$;	• 统计量: $t_{\mathrm{obs}} = \dfrac{\hat{\beta}_1}{\hat{\sigma}(\hat{\beta}_1)}$;

其中 $\hat{\sigma}(\hat{\beta}_1) = \sqrt{\dfrac{S_e}{S_{xx}(n-2)}}$.

- 拒绝域 $|t_{obs}| \geqslant t_{1-\alpha/2}(n-2)$；
- 正半轴拒绝域：$R.R^{+}$：$t_{obs} \geqslant t_{1-\alpha}(n-2)$；（或负半轴拒绝域：$R.R^{-}$：$t_{obs} \leqslant -t_{1-\alpha}(n-2)$；）
- p 值：$2P(t \geqslant |t_{obs}|)$.
- 正半轴 p 值：$P(t \geqslant t_{obs})$.（或负半轴 p 值：$P(t \leqslant t_{obs})$.）

β_1 的置信度为 $1-\alpha$ 的置信区间为

$$\hat{\beta}_1 \pm t_{1-\alpha/2}(n-2)\,\hat{\sigma}(\hat{\beta}_1) \equiv \hat{\beta}_1 \pm t_{1-\alpha/2}(n-2)\,\frac{\sqrt{S_e}}{\sqrt{n-2}\,\sqrt{S_{xx}}}\ .$$

如果整个置信区间包含在正半轴内，则表示解释变量与应变量正相关；如果整个置信区间包含在负半轴内，则表示解释变量与应变量负相关；如果整个置信区间包含 0，则无法判别解释变量与应变量的相关性．基于置信区间的推断结论与双边检验的结论相同．

2. 关于 β_0 的统计推断

由式(1.9)和式(1.7)可知，

$$\sqrt{n-2}\,\frac{\hat{\beta}_0-\beta_0}{\sqrt{S_e\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}} \sim t(n-2)\ .$$

从而可知，β_0 的置信度为 $1-\alpha$ 的置信区间为

$$\beta_0 \pm t_{1-\alpha/2}(n-2)\,\frac{\sqrt{S_e\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}}{\sqrt{n-2}}\ .$$

而假设检验问题

$$H_0:\beta_0=0;\quad H_1:\beta_0 \neq 0$$

的拒绝域为

$$\sqrt{n-2}\,\frac{|\hat{\beta}_0|}{\sqrt{S_e\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}} \geqslant t_{1-\alpha/2}.$$

一般来说，对于 β_0 的统计推断并不十分重要，因为 β_0 与预测变量 X 的坐标选取有关．比如，对 X 的坐标变换 $X'=X+a$，不影响应变量 Y 与 X 之间的相关关系，即 $\hat{\beta}_1$ 不改变，但会改变 $\hat{\beta}_0$ 的值。比如，作变换 $X'=X-\bar{X}$ 之后，可以证明 $\hat{\beta}_0=0$.

3. 关于 σ^2 的统计推断

由式(1.7)的 $\frac{S_e}{\sigma^2} \sim \chi^2(n-2)$ 可知，σ^2 的置信度为 $1-\alpha$ 的置信区间为

$$\left[\frac{S_e}{\chi^2_{1-\alpha/2}(n-2)}\ ,\ \frac{S_e}{\chi^2_{\alpha/2}(n-2)}\right].$$

关于假设 H_0：$\sigma^2=\sigma_0^2$ 的显著性水平为 α 的拒绝域为

$$\frac{S_e}{\sigma_0^2} \geqslant \chi^2_{1-\alpha/2}(n-2) \text{ 或 } \frac{S_e}{\sigma_0^2} \leqslant \chi^2_{\alpha/2}(n-2).$$

4. 关于估计值的统计推断

得到回归方程的参数估计后，通常有两个目的，一是研究应变量 Y 与预测变量 X 之

间的相关关系；二是对给定预测变量 X 的取值 x_0，使用这个模型来估计应变量的取值 Y_0.

回归方程

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

中的 $\hat{Y}_0$ 是给定 x_0 时 Y_0 的数学期望 $E[Y_0]$ 的无偏估计，即

$$E[\hat{Y}_0] = E[\hat{\beta}_0] + E[\hat{\beta}_1] x_0 = \beta_0 + \beta_1 x_0.$$

$$\begin{aligned}\mathrm{Var}[\hat{Y}_0] &= \mathrm{Var}[\hat{\beta}_0] + x_0^2 \mathrm{Var}[\hat{\beta}_1] + 2x_0 \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\ &= \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right)\sigma^2 + \frac{x_0^2}{S_{xx}}\sigma^2 - 2x_0 \frac{\bar{X}}{S_{xx}}\sigma^2 \\ &= \left(\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{S_{xx}}\right)\sigma^2 .\end{aligned}$$

由于 σ^2 未知，用 $\frac{S_e}{n-2}$ 代替，得到 $E[Y_0]$ 的区间估计

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{1-\alpha/2}(n-2) \sqrt{\frac{S_e}{n-2}\left(\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{S_{xx}}\right)} .$$

由于数学期望 $E[Y_0]$ 是在 x_0 处做多次观测的平均值，如果仅考虑一次性预测，还需考虑回归方程的误差，即 $Y_0 - E[Y_0] = Y_0 - \hat{Y}_0$. 显然，该误差的均值为 0，方差为

$$\mathrm{Var}[Y_0 - \hat{Y}_0] = \mathrm{Var}[Y_0] + \mathrm{Var}[\hat{Y}_0] + 2\mathrm{Cov}(Y_0, \hat{Y}_0) .$$

因为 $\hat{Y}_0$ 为 $(Y_1, Y_2, \cdots, Y_n)$ 的线性组合，而 Y_0 为新点 x_0 处的对应值，所以 Y_0 与 $\hat{Y}_0$ 相互独立，进而 $\mathrm{Cov}(Y_0, \hat{Y}_0) = 0$. 因此，

$$\mathrm{Var}[Y_0 - \hat{Y}_0] = \mathrm{Var}[Y_0] + \mathrm{Var}[\hat{Y}_0] = \sigma^2 + \left(\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{S_{xx}}\right)\sigma^2 .$$

由此可得新点 x_0 处对应值的置信度为 $(1-\alpha)$ 的区间估计为

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{1-\alpha/2}(n-2) \sqrt{\frac{S_e}{n-2}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{S_{xx}}\right)} .$$

各种统计软件都给出这两种置信区间，一种是对预测值的数学期望的区间估计，另一种是对单独观测点的预测值的区间估计. 两种估计的点估计相同，置信区间的长短不同.

5. 关于相关系数的统计推断

我们知道，相关系数度量两个变量之间（线性）相关的程度，相关系数的正负号与回归直线的斜率的正负号相同，解释变量 X 或应变量 Y 的线性变换不改变相关系数的值，相关系数既可以刻画两个独立变化的量之间的相关程度，也可以刻画相互依赖的量（如体重与身高）之间的相关程度.

一元线性回归分析中，解释变量 X 与应变量 Y 之间的相关系数 r 的估计量为皮尔逊（Pearson）相关系数，

$$\hat{r}_{xy} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}, \quad -1 \leqslant \hat{r}_{xy} \leqslant 1 .$$

$\hat{r}_{xy}$的标准差为 $\mathrm{SE}=\sqrt{\dfrac{1-r^2}{n-2}}$. 借助 Pearson 相关系数临界值表，可对相关系数作假设检验和区间估计.

6. 一元线性回归分析中的方差分析

方差分析的想法是将应变量 Y 的全变差分解为解释变量 X 导致的变差与随机因素导致的变差之和.

$$\begin{aligned}\mathrm{SST}&=\sum_{i=1}^{n}(Y_i-\bar{Y})^2=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2+\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2\\&=\mathrm{SSE}+\mathrm{SSR}.\end{aligned}$$

式中，总平方和(SST)的自由度为 $\mathrm{df}_{\mathrm{total}}=n-1$；残差平方和(SSE)的自由度为 $\mathrm{df}_{\mathrm{error}}=n-2$；回归平方和(SSR)的自由度为 $\mathrm{df}_{\mathrm{model}}=1$.

由于 $\beta_1=0$ 时，$F=\dfrac{\mathrm{SSR}}{\mathrm{SSE}/(n-2)}\sim F(1,\ n-2)$，因此，对于原假设为 $\beta_1=0$ 的显著性检验的拒绝域为 $F\geqslant F_{1-\alpha}(1,\ n-2)$.

另外，定义

$$R^2=\frac{\mathrm{SSR}}{\mathrm{SST}},$$

则 $0\leqslant R^2\leqslant 1$.

如果 $R^2=0$，即 $\mathrm{SST}=\mathrm{SSE}$，此时 X 与 Y 独立，也就是 Y 的变化与 X 无关，不可能用 X 的变化来解释 Y 的变化，此时也有 $\beta_1=0$.

如果 $R^2=1$，即 $\mathrm{SST}=\mathrm{SSR}$，此时 X 与 Y 存在完全的线性关系，也就是 Y 的取值完全由 X 的取值确定.

如果 R^2 较小，只说明 X 与 Y 之间的线性关系不成立，但 X 还有可能影响到 Y，只是它们之间的关系不是线性的. 在一元回归模型中，R^2 为 Pearson 相关系数的平方.

1.2 多元线性回归及相关性

虽然单变量回归模型很常用，但是，研究人员通常考虑多因素对应变量的影响.

多元线性回归分析理论模型假设从属变量(即响应变量或应变量)Y 与回归变量(即预测变量或自变量)X_1，X_2，…，X_m 存在不精确的线性依赖关系，需添加误差项 ε，并由 ε 的分布特性来刻画除 X_1，X_2，…，X_m 之外的其它因素对 Y 的影响，所以多元线性回归的模型为

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_mX_m+\varepsilon. \tag{1.10}$$

回归系数 β_k($k=1,\ 2,\ \cdots,\ m$)表示在其它回归变量(即预测变量)的取值给定的条件下，第 k 个回归变量增加一个单位时，对从属变量平均响应的影响效果. 回归系数 β_k 还可以解释为

$$\beta_k=\frac{\partial E[Y\mid X_1,X_2,\cdots,X_m]}{\partial X_k},\quad k=1,2,\cdots,m.$$

也就是说，β_k 是当保持其它独立变量不变的条件下，X_k 有单位改变量时，响应变量 Y 的

平均改变量.

回归变量通常也称为独立变量、外生变量、协变量、输入变量以及预测变量等等. 应变量与多个独立变量的非线性关系一般地表示为

$$Y = g(X_1, X_2, \cdots, X_n).$$

1.2.1 多元线性回归模型的参数估计

设$(x_{i1}, x_{i2}, \cdots, x_{im}; y_i)(i=1, 2, \cdots, n)$为解释变量$X_1, X_2, \cdots, X_m$与应变量$Y$的$n$组观测值. 记

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

则模型(1.10)可写为

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}. \tag{1.11}$$

而$(y_1, y_2, \cdots, y_n)^{\mathrm{T}}$为$\boldsymbol{Y}$的观测值. 通常模型(1.10)中都假设$\boldsymbol{\varepsilon} \sim N(0, \sigma^2\boldsymbol{I})$, 亦即

$$\boldsymbol{Y} \sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I}). \tag{1.12}$$

模型的残差平方和为

$$\mathrm{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = (\boldsymbol{Y} - \hat{\boldsymbol{Y}})^{\mathrm{T}}(\boldsymbol{Y} - \hat{\boldsymbol{Y}}),$$

式中, $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_m x_{im}$, $i=1, 2, \cdots, n$. 参数的最小二乘估计就是选取$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_m$, 使残差平方和SSE达到最小, 通过令SSE关于$\beta_0, \beta_1, \cdots, \beta_m$的一阶偏导数为0, 得到

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X\beta} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}.$$

若$\boldsymbol{X}$列满秩, 则$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$可逆, 从而$\boldsymbol{\beta}$的最小二乘估计量为

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{Y}. \tag{1.13}$$

由式(1.12)可知,

$$\begin{aligned} E[\hat{\boldsymbol{\beta}}] &= E[(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}] \\ &= (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}E[\boldsymbol{Y}] \\ &= (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X\beta} \\ &= \boldsymbol{\beta}. \end{aligned}$$

$$\begin{aligned} \mathrm{Var}[\hat{\boldsymbol{\beta}}] &= \mathrm{Cov}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}) \\ &= \mathrm{Cov}((\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}, (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}) \\ &= (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\mathrm{Cov}(\boldsymbol{Y}, \boldsymbol{Y})\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}. \end{aligned}$$

从而

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}). \tag{1.14}$$

另外，记 $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$，则

$$\begin{aligned}
\mathrm{SSE} &= (\boldsymbol{Y} - \hat{\boldsymbol{Y}})^{\mathrm{T}}(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) \\
&= (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\
&= (\boldsymbol{Y} - \boldsymbol{P}\boldsymbol{Y})^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{P}\boldsymbol{Y}) \\
&= \boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P})^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y} \\
&= \boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}.
\end{aligned}$$

σ^2 的无偏估计为

$$\widehat{\sigma^2} = \frac{\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}}{n - m - 1}.$$

所有的统计软件包都能计算出参数的最小二乘估计和标准差的估计.

1.2.2 多元线性回归模型的假设检验

1. 模型的整体检验——F 检验

检验的目的是判断每个回归变量是否都与应变量有线性关系.

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_m = 0; \quad H_A: \beta_i (i = 1, 2, \cdots, m) \text{ 不全为 } 0.$$

统计量为

$$F_{\mathrm{obs}} = \frac{\mathrm{MSR}}{\mathrm{MSE}} = \frac{R^2/m}{(1 - R^2)/(n - m - 1)}.$$

拒绝域为

$$F_{\mathrm{obs}} \geqslant F_{1-\alpha}(m, n - m - 1).$$

p 值为 $2P(F \geqslant F_{\mathrm{obs}})$.

2. 个别系数的检验——t 检验

检验的目的是在控制其它解释变量影响的前提下，判断应变量是否与某一个解释变量相关.

$$H_0: \beta_j = 0; \quad H_A: \beta_j \neq 0 (\text{双边检验}).$$

统计量为

$$t_{\mathrm{obs}} = \frac{\hat{\beta}_j / \hat{\sigma}(\hat{\beta}_j)}{\sqrt{\dfrac{\mathrm{SSE}}{n - m - 1}}} \sim t(n - m - 1).$$

其中，$\hat{\sigma}(\hat{\beta}_j)$ 为 $\hat{\beta}_j$ 的标准差，它为 $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$ 对角线上 β_j 相应位置的值的平方根.

拒绝域为

$$|t_{\mathrm{obs}}| \geqslant t_{1-\alpha/2}(n - m - 1).$$

p 值为 $2P(t \geqslant |t_{\mathrm{obs}}|)$.

3. 系数线性组合的检验——F 检验

检验的目的是在控制某些解释变量影响的前提下，判断应变量是否与其它解释变量的线性组合相关.

设 $\boldsymbol{L}$ 为 m 列 $k(k\leqslant m)$ 行的矩阵，秩为 s.

$$H_0:\boldsymbol{L\beta}=\boldsymbol{c};\quad H_A:\boldsymbol{L\beta}\neq\boldsymbol{c}(\text{双边检验}).$$

统计量为

$$F_{\text{obs}}=\frac{(\boldsymbol{L}\hat{\boldsymbol{\beta}}-\boldsymbol{c})^{\mathrm{T}}(\boldsymbol{L}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{L}^{\mathrm{T}})^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}}-\boldsymbol{c})/s}{\dfrac{\text{SSE}}{n-m-1}}\sim F(s,n-m-1).$$

拒绝域为

$$F_{\text{obs}}\geqslant F_{1-\alpha}(s,n-m-1).$$

p 值为 $2P(F\geqslant F_{\text{obs}})$.

值得强调的是，在实际应用中，有些模型的解释变量既有数值型数据，又有分类型数据(比如性别)，此时需要建立含亚元的回归模型.

如果分类变量有 k 个水平，则需有 $k-1$ 个亚元变量，第 j 个亚元变量只在第 j 个水平状态下取值为 1，其它水平状态下取值为 0. 而分类变量的基准水平是所有 $k-1$ 个亚元取值为 0. 对应于亚元的回归系数表示在所有数值变量保持不变时，该亚元的取值水平与基准水平的平均差距.

有关模型的解释变量既有数值型数据又有分类型数据的回归分析，将在第 4 章详细讨论.

1.3 线性回归的应用

线性回归被广泛应用于生物学、行为科学和社会科学等诸多领域，用来刻画变量之间的关系，是这些领域研究中最重要的工具之一.

1. 趋势线

趋势线是在考虑了其它一些构成要素之后，反映数据随时间变化的一种长期运动趋势. 它探索一个特殊的数据集(如 GDP、石油价格或股票价格等等)在一个时期内是否有递增或递减. 趋势线可以通过肉眼观察数据的散点图而轻易得到，但是，其位置和斜率须经诸如线性回归这样的统计技术才能得到. 尽管有时根据散点图反映的弯曲程度可用高阶多项式作趋势线，通常趋势线为直线.

趋势线有时用于商业分析，用其来反映一种事物随时间的变化，此时其优势在于简单. 趋势线往往用来讨论一个特殊行动或事件引起的某些结果随时间的变化(例如，训练或广告效果比较). 作趋势线是一种简便易行的方法，它不需要对照组数据、试验设计和复杂的分析技巧，但其缺陷是当有其它潜在变化影响数据时，有效性较差.

2. 流行病学

较早应用回归分析的典型例子是探求吸烟与死亡率、患病率的关系，研究者通常在其回归分析模型中引入多个变量并力争剔除那些可能产生虚假相关性的变量. 例如，就吸烟的情形，研究者可能会将社会经济状态因素包含在模型中，以保证观测到吸烟对死亡率的影响与受教育程度和收入等无关. 但是，使用回归分析时又不可能将所有的、可能混杂的因素都包含在模型中. 例如，可能某种基因会导致死亡率提高或使吸烟者吸更多烟. 基于