第一章 引论

Tianxiao Pang

Zhejiang University

September 10, 2019

内容

内容

内容

本课程(回归分析)主要讨论线性(统计)模型. 线性模型是现代统计学中应用最为广泛的模型之一, 而且也是其它统计模型研究或应用的基础. 理由如下:

- 现实世界中, 许多变量之间具有线性或近似线性的依赖关系;
- 现实世界中, 虽然有些变量之间的关系是非线性, 但往往可以通过适当的变换, 使得新变量之间具有线性或者近似线性的关系;
- 线性关系是数学中最基本的关系, 容易处理.

线性回归模型

现实生活中变量与变量之间的关系:

(1) 确定性关系: 变量之间的关系可用数学函数来表示. 例如, 一物体做自由落体时, 时间t与下落高度s这两个变量之间的关系可用$s = gt^2/2$来表示.

(2) 相关关系: 变量之间的关系不能用数学函数来刻画, 但具有一定的"趋势性"关系. 例如, 人的身高x与体重y这两个变量, 它们之间不具有确定性关系, 但人的身高越高, 往往体重也越重. 人的身高与体重具有相关关系. 父亲身高x与儿子身高y之间也具有相关关系.

回归分析的研究对象是具有相关关系的变量, 研究目的是寻求它们之间客观存在的依赖关系.

在以上例子中, y通常称为因变量(dependent variable)或者响应变量(response variable), x称为自变量(independent variable), 或解释变量(explanatory variable), 或协变量(covariate), 或预报变量(predictor variable). y 的值可看成由两部分组成: 由x决定的部分(记为$f(x)$) 以及其它未加考虑的因素所产生的影响, 后者被称为随机误差, 记作e. 因此, 自然地我们有下列模型:

$$y = f(x) + e.$$

特别地, 若$f(x) = \beta_0 + \beta_1 x$, 则

$$y = \beta_0 + \beta_1 x + e. \tag{1.1.1}$$

我们称上式为一元线性回归模型, 称β_0为回归常数, 称β_1为回归系数. 有时, 我们把β_0和β_1统称为回归系数.

记$(x_i, y_i), i = 1, \cdots, n$为来自$(x, y)$的样本. 若(1.1.1)成立, 则$(x_i, y_i)$应满足关系式

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \cdots, n. \tag{1.1.2}$$

基于以上样本信息, 应用适当的统计方法, 可得到β_0和β_1的点估计$\hat{\beta}_0$和$\hat{\beta}_1$. 我们称

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{1.1.3}$$

为(经验)回归直线或(经验)回归方程.

注: 通常假设随机误差$\{e_i, i \geq 1\}$是i.i.d.序列, 且与$\{x_i, i \geq 1\}$独立, 同时满足$\mathsf{E}(e_i) = 0$. 因此, (1.1.3)其实是回归函数

$$\mathsf{E}(y|x) = \beta_0 + \beta_1 x$$

的一个估计. 回归分析的首要问题是估计回归函数.

“回归”(regression)一词的由来:

英国统计学家Galton为了研究父代与子代身高的关系, 收集了1078对父亲及儿子身高的数据. 以x表示父亲身高, y表示儿子身高, 将1078对数据(x_i, y_i)画在直角坐标图纸上, 他发现散点图大致呈直线形状. 即总的趋势是, 当父亲身高增加时, 儿子的身高也倾向于增加. 但经过进一步的分析, Galton发现了一个有趣的现象—-回归效应.

计算得到$\bar{x} = 68$(单位: 英寸, 下同), $\bar{y} = 69$. 即子代身高平均增加了1英寸. 这样, 若父亲身高为x, 则他儿子的平均身高大致应为$x + 1$. 但Galton发现, 当父亲身高为72时, 他们儿子的平均身高仅为71; 而当父亲身高为64时, 他们的儿子的平均身高为67.

Galton认为: 大自然具有一种约束力, 使人类身高的分布在一定时期内相对稳定而不产生两极分化, 这就是所谓的回归效应.

例1.1.1 一个公司的商品销售量与其广告费有密切关系, 一般说来在其它因素(如产品质量等)保持不变的情况下, 它用在广告上的费用愈高, 商品销售量也会越多. 因此, 广告费x与销售量y是一种相关关系. 为了进一步研究这种关系, 根据过去的记录$(x_i, y_i), i = 1, \cdots, n$, 采用线性回归模型(1.1.2), 假定计算出$\hat{\beta}_0 = 1608.5$, $\hat{\beta}_1 = 20.1$, 于是得到回归方程

$$\hat{y} = 1608.5 + 20.1x.$$

这告诉我们: 广告费每增加一个单位, 该公司的销售量就平均(或大约)增加20.1个单位.

在实际问题中, 影响因变量的主要因素往往很多, 这就需要考虑含多个自变量的回归问题.

假设因变量y和p个自变量满足如下的多元线性回归模型:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + e. \tag{1.1.4}$$

若$(x_{i1}, \cdots, x_{ip}, y_i), i = 1, \cdots, n$为相应的样本, 则它们满足关系式

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + e_i, \quad i = 1, \cdots, n. \tag{1.1.5}$$

注: 线性模型指的是它关于未知参数$\beta_0, \beta_1, \cdots, \beta_p$是线性的.

引入矩阵符号:

$$\boldsymbol{Y}=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{X}=\begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \boldsymbol{\beta}=\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \boldsymbol{e}=\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

则(1.1.5)可简写为

$$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{e}. \tag{1.1.6}$$

注: 通常, 称$\boldsymbol{X}$为设计矩阵.

关于随机误差向量$\boldsymbol{e}$, 通常假定它满足Gauss-Markov假设:

- 均值为零, 即$\mathsf{E}(e_i) = 0$;
- 方差齐性, 即$\mathsf{Var}(e_i) = \sigma^2$;
- 彼此不相关, 即$\mathsf{Cov}(e_i, e_j) = 0,\ i \neq j$.

这三条假设分别要求: 误差项不包含任何系统的趋势;每一个y_i在其均值附近波动程度是一致的(假定自变量不是随机的); 不同次的观测(即$y_i, i = 1, \cdots, n$)是不相关的.

假设$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p)'$为$\boldsymbol{\beta}$的一个估计, 则可得到(经验)回归方程

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p \tag{1.1.7}$$

或者写成

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}. \tag{1.1.8}$$

例1.1.2 在经济学中, 著名的柯布-道格拉斯(Cobb-Douglas)生产函数为

$$Q_t = aL_t^b K_t^c,$$

这里, Q_t, L_t, K_t分别表示第t年的产值、劳动投入量和资金投入量. a, b, c为参数. 为了估计a, b, c, 对Cobb-Douglas生产函数取自然对数, 得

$$\ln(Q_t) = \ln a + b\ln(L_t) + c\ln(K_t).$$

再令

$$y_t = \ln(Q_t), x_{t1} = \ln(L_t), x_{t2} = \ln(K_t), \beta_0 = \ln a, \beta_1 = b, \beta_2 = c.$$

则问题转化为在下列的线性回归模型中估计未知参数a, b, c:

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + e_t, \quad t = 1, \cdots, T.$$

方差分析模型

上一节的线性回归模型中, 自变量一般都是连续变量, 研究的目的是寻求因变量和自变量之间客观存在的依赖关系. 但有时, 自变量是示性变量, 这种变量往往表示某种效应的存在与否, 因而只能取0和1两个值. 这种模型是比较两个或者多个因素的效应大小的一种有力工具. 人们称这种模型为方差分析模型.

例1.2.1 某农业科学研究机构欲比较三种小麦品种的优劣, 安排了一种比较试验. 为保证试验结果的客观性, 他们选择了六块面积相等, 土质肥沃程度一样的田地, 每一种小麦播种在其中的两块田内, 并给以几乎完全相同的田间管理. 用y_{ij}表示种第i种小麦的第j块田的产量, 那么可认为

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1, 2, 3; \ j = 1, 2.$$

这里, μ表示总平均值, α_i表示种第i种小麦品种的效应, e_{ij}是随机误差, 它表示所有其它未加控制因素以及各种误差的总效应.

若采用矩阵符号, 则上述模型可改写为

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \end{pmatrix}$$

或

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}.$$

它与上一节的线性回归模型的矩阵形式(1.1.6)完全一样. 不同的是, 在(1.1.6)中, 除第一列外, 设计矩阵$\boldsymbol{X}$的元素原则上可以取任意连续值, 而在现在的模型中, 设计矩阵$\boldsymbol{X}$的所有元素只能取0和1两个值. 我们对方差分析模型进行统计分析的目的是比较这三种小麦品种, 即比较它们的效应$\alpha_1, \alpha_2, \alpha_3$的大小.

应用矩阵符号, 可以看出线性回归模型和方差分析模型具有相同的形式. 但它们的用处却有较大差别, 线性回归模型本质上用于描述变量之间的依赖关系, 方差分析模型主要用于比较效应的大小.

还有其它一些更加复杂的模型, 如协方差分析模型(自变量中既有连续变量又有示性变量)等, 也与线性回归模型有较大的联系, 可以采用线性回归分析的方法进行研究. 略.

应用概述

对回归模型进行的统计分析, 通常称为回归分析(Regression Analysis). 应用:

1. 描述变量之间的关系. 根据因变量和自变量的观测值, 通过一些统计推断方法, 我们可以建立起因变量和自变量之间的(经验)回归方程. 这个方程用来刻画因变量和自变量之间的相依关系. 但需注意的是, 获得回归方程后, 我们还需考虑这个回归方程是否真正刻画了因变量和自变量之间客观存在的依赖关系. 这是因为, 当我们应用线性回归模型对数据进行分析时, 面临着模型选择, 自变量选择, 误差假设适用性等问题. 处理不当, 结果会呈现一定的误差. 因此, 在实际中, 建立回归方程的过程是一个迭代的过程. 先选择一个初始模型, 基于数据得到回归方程后, 经过一些统计检验后结合专业知识的分析, 若认为初始模型不够合理, 则对其进行适当修正, 或改变估计方法, 然后重新建立回归方程. 重复上述过程, 直到所得到的回归方程从诸多角度考察都比较满意为止.

2. 分析变量之间的相互关系. 假设我们已得到一个比较满意的回归方程:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p. \tag{1.3.1}$$

适当消除自变量$x_1, \cdots, x_p$量纲的影响后, 回归系数β_i的估计$\hat{\beta}_i$的大小在一定程度上反映了自变量x_i对因变量y的影响大小.
当$\hat{\beta}_i > 0$时, y与x_i是正相关关系; 当$\hat{\beta}_i < 0$时, y与x_i是负相关关系; $|\hat{\beta}_i|$越大, 表明x_i这个自变量越重要.

3. 预测. 得到一个满意的回归方程(1.3.1)后, 对于自变量的一组特定值$(x_{01}, \cdots, x_{0p})$, 我们可以得到因变量的预测值

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_p x_{0p}.$$