第一章 引论

本课程主要讨论线性模型. 线性模型利用自变量的线性组合去解释因变量. 为何重点关注线性模型? 这是因为: (1) 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. (2) 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函数变换, 使得新变量之间具有线性或者近似线性的关系. (3) 线性关系是数学中最基本的关系, 容易处理.

线性模型非常灵活, 它广泛应用于物理学、工程、社会科学和商业等领域. 线性模型是现代统计学中应用最广泛的模型之一, 也是其他统计模型研究或应用的基础.

1.1 线性回归模型

在现实生活中, 变量与变量之间存在着以下的两种关系:

(1) 确定性关系. 即变量之间的关系可用数学函数来刻画. 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. 即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, 他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(或母亲)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. 在以上例子中, 目标变量y通常称被为因变量(dependent variable)或者响应变量(response variable), 用来解释或预测y的变量x被称为自变量(independent variable), 或解释变量(explanatory variable), 或协变量(covariate), 或预报变量(predictor variable). y的取值可看成由两部分组成: 由x决定的部分(记为$f(x)$)以及其他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, 有下列模型:

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

特别地, 若$f(x) = \beta_0 + \beta_1 x$, 则

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

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

设$\{(x_i, y_i), i = 1, \cdots, n\}$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, 得到β_0和β_1的点估计$\hat{\beta}_0$和$\hat{\beta}_1$. 称

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

为(经验)回归方程或(经验)回归直线.

通常假设随机误差$\{e_i, i \geq 1\}$是独立同分布(i.i.d.)随机变量序列, 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. 回归函数刻画了因变量的平均大小与自变量的相依关系.

回归分析有两个主要目标: (1) 揭示因变量(在平均意义下)与自变量之间的依赖关系. (2) 对因变量的将来值或无法观测的值进行预测. 对于第一个目标, 通常等价于估计回归函数, 在线性模型框架下等价于估计回归系数. 得到一个理想的回归方程后, 很自然地, 可完成第二个目标. 在处理数据前, 需要清楚我们的目标是什么, 目标不同, 回归分析的结果往往也会有所不同.

回归模型通常来自哪里? (1) 来自物理理论. 例如, 胡克定律说弹簧的伸长量与附加的重量成正比. 这样的模型通常出现在物理科学和工程领域. (2) 来自对以前数据的处理经验. 若过去使用的类似数据是以某种方式建模的, 那么很自然地, 我们会把相同的模型应用于当前的数据. 这样的模型通常出现在社会科学领域. (3) 之前不存在任何经验或想法, 该模型来自对数据分析的探索.

直接源自物理理论的模型相对不常见, 因此线性模型通常只能被视为对复杂现实的近似. 我们希望它能很好地解释因变量与自变量之间的相关关系或进行因变量的预测. 一个好的模型就像一张地图, 能指引我们完成统计分析的目的. 若无特别说明, 假设所讨论的模型是正确的.

例1.1.1 一个公司的商品销售量与其广告费有密切关系, 一般说来在其他因素(如产品质量等)保持不变的情况下, 它用在广告上的费用越高, 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:

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

该回归方程告诉我们: 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有很多, 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$x_1, \cdots, x_p$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, 则他们满足关系式

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

注: 线性模型指的是因变量关于未知参数$\beta_0, \beta_1, \cdots, \beta_p$是线性的.

引入矩阵符号:

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

则(1.1.5)可简写为

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

通常, 称$\boldsymbol{X}$为设计矩阵或数据矩阵.

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

(1) 均值为零: $\mathsf{E}(e_i) = 0$.

(2) 方差齐性: $\mathsf{Var}(e_i) = \sigma^2$.

(3) 彼此不相关: $\mathsf{Cov}(e_i, e_j) = 0,\ i \neq j$.

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在其均值附近的波动程度是一致的(假定自变量不是随机的); 不同次的观测(即$y_i, y_j, i \neq j$)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

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

对模型(1.1.6)可以作如下的理解:

$$\begin{aligned} \boldsymbol{Y} &= \quad \boldsymbol{X}\boldsymbol{\beta} \quad + \quad \boldsymbol{e}, \\ \text{数据} &= \text{系统部分} + \text{随机部分}, \\ n\text{维} &= (p+1)\text{维} + (n-p-1)\text{维}, \end{aligned}$$

称$(n-p-1)$为模型的自由度(degrees of freedom). 当$n = p+1$时, 我们也许可以估计出所有的回归系数, 但没有多余的自由度可以估计其他参数(例如σ^2)或者做假设检验等统计推断, 此时称模型是饱和的. 当$n < p+1$时, 称模型是超饱和的, 此时, 若对模型不施加一些约束条件, 将无法估计出回归系数. 在本课程中, 若无特别说明, 默认$n > p+1$.

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

$$Q_t = a L_t^b K_t^c,$$

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, 对柯布-道格拉斯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

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

然后, 可以把问题看成是在下列的线性回归模型中估计未知参数a, b, c,

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

在这一节的最后, 来介绍一下“回归”(regression)一词的由来. “回归”, 这个乍看起来有点奇怪的词, 是由英国生物学家兼统计学家高尔顿(Francis Galton, 1822–1911; 达尔文的表弟)在1875年提出的. 当时, 他收集了934组(实际上是963组, 但其中的29组数据是非数值型的, 故没有使用它)父母亲身高和孩子身高的数据(该数据可从R软件的HistData这个package里找到, 数据集的名字为GaltonFamilies, 身高的量纲为英寸), 想研究父母身高和孩子身高的相关关系. 高尔顿使用childHeight作为因变量, midparentHeight作为自变量, 它等于(父亲身高+1.08×母亲身高)/2, 将它们的数据画在直角坐标图纸上(见图1.1.1), 发现散点图大致呈直线形状. 即总的趋势是, 当父母身高越高(越矮)时, 孩子的身高也倾向于越高(越矮). 因此, 可用一元线性回归模型

$$\text{childHeight} = \beta_0 + \beta_1 \text{midparentHeight} + e \tag{1.1.9}$$

来对childHeight和midparentHeight这两个变量做进一步的分析. 应用当时的估计方法(最小二乘法, 将在第三章介绍), 可得到

$$\hat{\beta}_0 = 22.636, \quad \hat{\beta}_1 = 0.637.$$

相应的回归直线见图1.1.2中的实线.

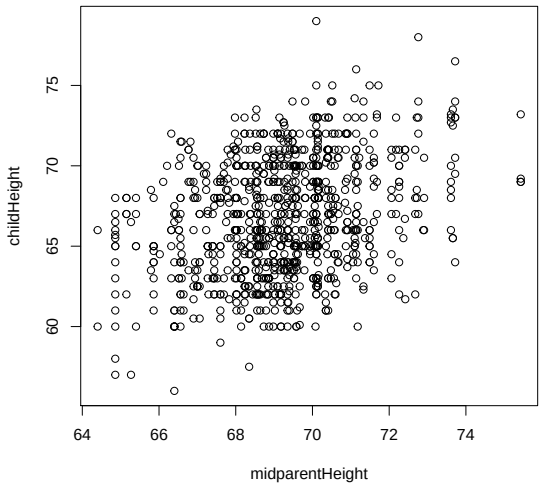

图1.1.1 散点图

需要说明的是, 用下面的等式进行建模, 与用回归模型(1.1.9)进行建模是等价的,

$$\frac{y - \bar{y}}{\text{sd}(y)} = r \frac{x - \bar{x}}{\text{sd}(x)}, \tag{1.1.10}$$

其中y表示childHeight, x表示midparentHeight, $\bar{y}$和$\bar{x}$表示样本均值, sd表示样本标准差, r表示x和y的样本相关系数. 这个等式是说, 因变量的一个标准变化等于

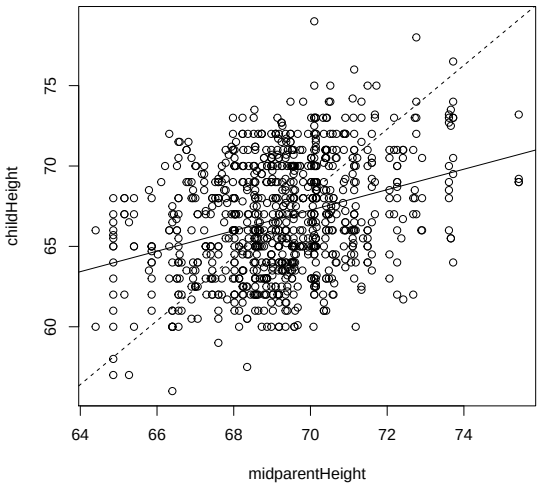

图1.1.2 回归直线(实线)

自变量的一个标准变化与相关系数的乘积. 把上述等式改写成

$$y = \left(\bar{y} - \frac{\mathrm{sd}(y)}{\mathrm{sd}(x)} r\bar{x}\right) + \frac{\mathrm{sd}(y)}{\mathrm{sd}(x)} r \cdot x, \tag{1.1.11}$$

可计算得到

$$\bar{y} - \frac{\mathrm{sd}(y)}{\mathrm{sd}(x)} r\bar{x} = 22.636, \quad \frac{\mathrm{sd}(y)}{\mathrm{sd}(x)} r = 0.637.$$

我们可能会天真地猜测: 父母身高的一个标准变化, 会导致孩子身高的(大约)一个标准变化. 这种猜测对应于(1.1.10)中$r = 1$或$r \approx 1$的情形. 当$r = 1$时, (1.1.11)中的截距和斜率分别是

$$\bar{y} - \frac{\mathrm{sd}(y)}{\mathrm{sd}(x)} r\bar{x} = -70.689, \quad \frac{\mathrm{sd}(y)}{\mathrm{sd}(x)} r = 1.986,$$

相应的直线见图1.1.2中的虚线. 图中的直线都经过$(\bar{x}, \bar{y})$这一点, 即他们的交点是$(\bar{x}, \bar{y})$. 从这张图上可以看出: 在现实社会中, 父母身高较高的, 他们的孩子的身高倾向于高于平均水平, 但并不像虚线(虚线代表我们的朴素猜测)所示的那么高; 类似地, 父母身高较矮的, 他们的孩子的身高会倾向于低于平均水平, 但并不像虚线所示的那么矮. 高尔顿认为: 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. 因此, 他把这种现象称为“回归平庸现象”, 也被称为“回归均值现象”或“回归效应”.

1.2 方差分析模型

在上一节的线性回归模型中, 自变量是定量变量, 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或利用自变量对因变量进行预测. 但有时, 自变量是定性变量, 这种变量往往表示某种效应的存在与否(可采用$0,1,\cdots$等数字进行编码). 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, 让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}表示第i种小麦在第j块田的产量, 那么可认为

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

这里, μ表示总平均值, α_i表示第i个小麦品种的效应, e_{ij}是随机误差, 它表示所有未加控制的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.

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, 但两个模型的建模目的却有较大差别. 线性回归的建模目的是描述变量之间的依赖关系或进行因变量的预测, 而方差分析的建模目的是比较效应的大小.

还有其他一些更加复杂的模型, 如协方差分析模型(自变量中既有定量变量又有定性变量)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.

1.3 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). 它有如下一些应用:

(1) 描述变量之间的依赖关系. 根据因变量和自变量的观测值, 通过一些统计推断方法, 可以建立起因变量和自变量之间的回归方程. 这个方程刻画了因变量在平均意义下和自变量之间的依赖关系.

但需注意的是, 在实际问题当中, 得到回归方程后, 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, 这需要做进一步的统计分析. 这是因为, 在实际问题当中, 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. 若处理不当, 统计分析的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, 并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) 分析变量之间的关系. 假设已得到一个比较满意的回归方程(消除自变量$x_1, \cdots, x_p$量纲的影响):

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

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这个自变量对于y的解释越重要.

(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$(x_{01}, \cdots, x_{0p})$, 可以得到因变量的预测值:

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

作业

1. 对GaltonFamilies数据集中的父亲身高(变量名为father)与孩子身高进行分析, 看看是否存在“回归效应”? 若是对母亲身高(变量名为mother)与孩子身高进行分析呢?

2. 在现实生活中, 你觉得还有哪些现象也存在着“回归效应”?