線性回歸

维基百科，自由的百科全书

在统计学中，**线性回归（Linear regression）**是利用称为线性回归方程的最小平方函數对一个或多个自变量和因变量之间关系进行建模的一种回归分析。这种函数是一个或多个称为回归系数的模型参数的线性组合。只有一个自变量的情况称为简单回归，大于一个自变量情况的叫做多元回归。（这反过来又应当由多个相关的因变量预测的多元线性回归区别，而不是一个单一的标量变量。）

在线性回归中，数据使用线性预测函数来建模，并且未知的模型参数也是通过数据来估计。这些模型被叫做线性模型。最常用的线性回归建模是给定X值的y的条件均值是X的仿射函数。不太一般的情况，线性回归模型可以是一个中位数或一些其他的给定X的条件下y的条件分布的分位数作为X的线性函数表示。像所有形式的回归分析一样，线性回归也把焦点放在给定X值的y的条件概率分布，而不是X和y的联合概率分布（多元分析领域）。

线性回归是回归分析中第一种经过严格研究并在实际应用中广泛使用的类型。这是因为线性依赖于其未知参数的模型比非线性依赖于其未知参数的模型更容易拟合，而且产生的估计的统计特性也更容易确定。

线性回归有很多实际用途。分为以下两大类：

1. 如果目标是预测或者映射，线性回归可以用来对观测数据集的和的值拟合出一个预测模型。当完成这样一个模型以后，对于一个新增的X值，在没有给定与它相配对的y的情况下，可以用这个拟合过的模型预测出一个y值。
2. 给定一个变量y和一些变量X1,...,Xp，这些变量有可能与y相关，线性回归分析可以用来量化y与Xj之间相关性的强度，评估出与y不相关的Xj，并识别出哪些Xj的子集包含了关于y的冗余信息。

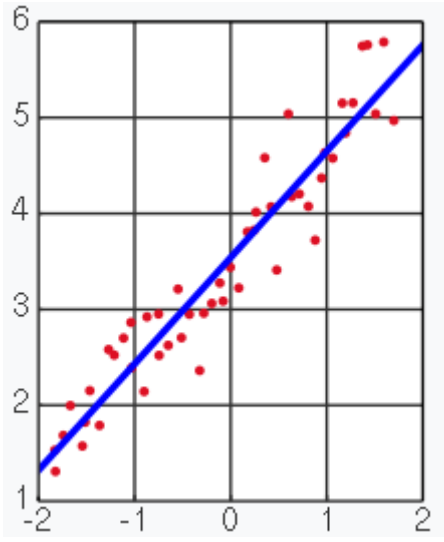
线性回归模型经常用最小二乘逼近来拟合，但他们也可能用别的方法来拟合，比如用最小化“拟合缺陷”在一些其他规范里（比如最小绝对误差回归），或者在桥回归中最小化最小二乘损失函数的惩罚。相反，最小二乘逼近可以用来拟合那些非线性的模型。因此，尽管“最小二乘法”和“线性模型”是紧密相连的，但他们是不能划等号的。

目录

簡介

理論模型

給一個隨機樣本 $(Y_i, X_{i1}, \ldots, X_{ip}), i = 1, \ldots, n$ ，一個線性回歸模型假設回歸子Y_i和回歸量$X_{i1}, \ldots, X_{ip}$之間的關係是除了X的影響以外，還有其他的變數存在。我們加入一個誤差項ε_i（也是一個隨機變量）來捕獲除了$X_{i1}, \ldots, X_{ip}$之外任何對Y_i的影響。所以一個多變量線性回歸模型表示為以下的形式：



帶有一個自變量的線性回歸

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n$$

其他的模型可能被認定成非線性模型。一個線性回歸模型不需要是自變量的線性函數。線性在這裡表示Y_i的條件均值在參數β裡是線性的。例如：模型$Y_i = \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$在$\beta_1$和$\beta_2$裡是線性的，但在$X_i^2$裡是非線性的，它是$X_i$的非線性函數。

數據和估計

區分隨機變量和這些變量的觀測值是很重要的。通常來說，觀測值或數據（以小寫字母表記）包括了n個值$(y_i, x_{i1}, \ldots, x_{ip}),\ i = 1, \ldots, n.$

我們有$p + 1$個參數$\beta_0, \ldots, \beta_p$需要決定，為了估計這些參數，使用矩陣表記是很有用的。

$$Y = X\beta + \varepsilon$$

其中Y是一個包括了觀測值$Y_1, \ldots, Y_n$的列向量，ε包括了未觀測的隨機成份$\varepsilon_1, \ldots, \varepsilon_n$以及回歸量的觀測值矩陣$X$：

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

X通常包括一個常數項。

如果X列之間存在線性相關，那麼參數向量β就不能以最小二乘法估計除非β被限制，比如要求它的一些元素之和為0。

古典假設

- **樣本**是在**母體**之中**隨機抽取**出來的。
- 因變量Y在實直線上是**連續的**，
- 殘差項是**獨立**且**相同**分佈的(iid)，也就是說，殘差是独立随机的，且服從**高斯分佈**。

這些假設意味著**殘差項不依賴自變量的值**，所以ε_i和自變量X（預測變量）之間是相互獨立的。

在這些假設下，建立一個顯示線性回歸作為條件預期模型的**簡單線性回歸**，可以表示為：

$$\mathbf{E}(Y_i \mid X_i = x_i) = \alpha + \beta x_i$$

最小二乘法分析

最小二乘法估計

回歸分析的最初目的是估計模型的參數以便達到對數據的最佳拟合。在決定一個最佳拟合的不同標準之中，最小二乘法是非常優越的。這種估計可以表示為：

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

回歸推論

對於每一個$i = 1, \ldots, n$，我們用σ^2代表誤差項ε的方差。一個無偏誤的估計是：

$$\hat{\sigma}^2 = \frac{S}{n-p},$$

其中$S := \sum_{i=1}^{n} \hat{\varepsilon}_i^2$是誤差平方和（殘差平方和）。估計值和實際值之間的關係是：

$$\hat{\sigma}^2 \cdot \frac{n-p}{\sigma^2} \sim \chi^2_{n-p}$$

其中χ^2_{n-p}服從卡方分佈，自由度是$n-p$

對普通方程的解可以寫為：

$$\hat{\boldsymbol{\beta}} = (\mathbf{X^T X})^{-1}\mathbf{X^T y}.$$

這表示估計項是因變量的線性組合。進一步地說，如果所觀察的誤差服從正態分佈。參數的估計值將服從聯合正態分佈。在當前的假設之下，估計的參數向量是精確分佈的。

$$\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$$

其中$N(\cdot)$表示多變量正態分佈。

參數估計值的標準差是：

$$\hat{\sigma}_j = \sqrt{\frac{S}{n-p}\left[(\mathbf{X^T X})^{-1}\right]_{jj}}.$$

參數β_j的$100(1-\alpha)\%$置信區間可以用以下式子來計算：

$$\hat{\beta}_j \pm t_{\frac{\alpha}{2}, n-p}\hat{\sigma}_j.$$

誤差項可以表示為：

$$\hat{\mathbf{r}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}(\mathbf{X^T X})^{-1}\mathbf{X^T y}.$$

單變量線性回歸

單變量線性回歸，又稱簡單線性回歸（simple linear regression, SLR），是最簡單但用途很廣的回歸模型。其回歸式為：

$$Y = \alpha + \beta X + \varepsilon$$

為了從一組樣本(y_i, x_i)（其中$i = 1, 2, \ldots, n$）之中估計最合適（誤差最小）的α和β，通常採用最小二乘法，其計算目標為最小化殘差平方和：

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

使用微分法求極值：將上式分別對α和β做一階偏微分，並令其等於0：

$$\begin{cases} n\,\alpha + \sum_{i=1}^{n} x_i\,\beta = \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i\,\alpha + \sum_{i=1}^{n} x_i^2\,\beta = \sum_{i=1}^{n} x_i y_i \end{cases}$$

此二元一次線性方程組可用克萊姆法則求解，得解$\hat{\alpha},\ \hat{\beta}$：

$$\hat{\beta} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \bar{y} - \bar{x}\hat{\beta}$$

$$S = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} y_i^2 - \frac{n(\sum_{i=1}^{n} x_i y_i)^2 + (\sum_{i=1}^{n} y_i)^2 \sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$\hat{\sigma}^2 = \frac{S}{n-2}.$$

協方差矩陣是：

$$\frac{1}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}$$

平均響應置信區間為：

$$y_d = (\alpha + \hat{\beta} x_d) \pm t_{\frac{\alpha}{2},n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_d - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

預報響應置信區間為：

$$y_d = (\alpha + \hat{\beta} x_d) \pm t_{\frac{\alpha}{2},n-2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_d - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

方差分析

在方差分析（ANOVA）中，總平方和分解為兩個或更多部分。

總平方和SST (sum of squares for total)是：

$$\mathrm{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \quad \text{，其中：} \quad \bar{y} = \frac{1}{n}\sum_{i} y_i$$

同等地：

$$\mathrm{SST} = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i} y_i\right)^2$$

回歸平方和SSReg (sum of squares for regression，也可寫做**模型平方和**，SSM，sum of squares for model)是：

$$\mathrm{SSReg} = \sum (\hat{y}_i - \bar{y})^2 = \hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{y} - \frac{1}{n} \left(\mathbf{y^T u u^T y} \right),$$

殘差平方和SSE (sum of squares for error)是：

$$\mathrm{SSE} = \sum_i (y_i - \hat{y}_i)^2 = \mathbf{y^T y} - \hat{\boldsymbol{\beta}}^\mathbf{T} \mathbf{X^T y}.$$

總平方和SST又可寫做SSReg和SSE的和：

$$\mathrm{SST} = \sum_i (y_i - \bar{y})^2 = \mathbf{y^T y} - \frac{1}{n} \left(\mathbf{y^T u u^T y} \right) = \mathrm{SSReg} + \mathrm{SSE}.$$

回歸係數R^2是：

$$R^2 = \frac{\mathrm{SSReg}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}.$$

其他方法

廣義最小二乘法

廣義最小二乘法可以用在當觀測誤差具有異方差或者自相關的情況下。

總體最小二乘法

總體最小二乘法用於當自變量有誤時。

廣義線性模式

廣義線性模式應用在當誤差分佈函數不是正態分佈時。比如指數分佈，伽瑪分佈，逆高斯分佈，泊松分佈，二項式分佈等。

穩健回歸

將平均絕對誤差最小化，不同於在線性回歸中是將均方誤差最小化。

線性回歸的應用

趨勢線

一條趨勢線代表著時間序列數據的長期走勢。它告訴我們一組特定數據（如GDP、石油價格和股票價格）是否在一段時期內增長或下降。雖然我們可以用肉眼觀察數據點在坐標系的位置大體畫出趨勢線，更恰當的方法是利用線性回歸計算出趨勢線的位置和斜率。

流行病学

有关吸烟对死亡率和发病率影响的早期证据来自采用了回归分析的观察性研究。为了在分析观测数据时减少伪相关，除最感兴趣的变量之外,通常研究人员还会在他们的回归模型里包括一些额外变量。例如，假设我们有一个回归模型，在这个回归模型中吸烟行为是我们最感兴趣的独立变量，其相关变量是经数年观察得到的吸烟者寿命。研究人员可能将社会经济地位当成一个额外的独立变量，已确保任何经观察所得的吸烟对寿命的影响不是由于教育或收入差异引起的。然而，我们不

可能把所有可能混淆结果的变量都加入到实证分析中。例如，某种不存在的基因可能会增加人死亡的几率，还会让人的吸烟量增加。因此，比起采用观察数据的回归分析得出的结论，随机对照试验常能产生更令人信服的因果关系证据。当可控实验不可行时，回归分析的衍生，如工具变量回归，可尝试用来估计观测数据的因果关系。

金融

資本資產定價模型利用線性回歸以及Beta係數的概念分析和計算投資的系統風險。這是從聯繫投資回報和所有風險性資產回報的模型Beta係數直接得出的。

经济学

线性回归是经济学的主要实证工具。例如，它是用来预测消费支出，固定投资支出，存货投资，一国出口产品的购买，进口支出，要求持有流动性资产，劳动力需求、劳动力供给。

参考文献

引用

来源

书籍

- Cohen, J., Cohen P., West, S.G., & Aiken, L.S *Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates. 2003.*
- Draper, N.R. and Smith, H. Applied Regression Analysis. Wiley Series in Probability and Statistics. 1998.
- Robert S. Pindyck and Daniel L. Rubinfeld. Chapter One. Econometric Models and Economic Forecasts. 1998.
- Charles Darwin. *The Variation of Animals and Plants under Domestication*. (1868) *(Chapter XIII describes what was known about reversion in Galton's time. Darwin uses the term "reversion".)*

刊物文章

- Galton, Francis. Regression Towards Mediocrity in Hereditary Stature (PDF). Journal of the Anthropological Institute. 1886, **15**: 246–263 [2008-12-30].

延伸阅读

- Pedhazur, Elazar J. Multiple regression in behavioral research: Explanation and prediction 2nd. New York: Holt, Rinehart and Winston. 1982. ISBN 0-03-041760-0.
- Barlow, Jesse L. Chapter 9: Numerical aspects of Solving Linear Least Squares Problems. (编) Rao, C.R. Computational Statistics. Handbook of Statistics **9**. North-Holland. 1993. ISBN 0-444-88096-8
- Björck, Åke. Numerical methods for least squares problems. Philadelphia: SIAM. 1996. ISBN 0-89871-360-9.
- Goodall, Colin R. Chapter 13: Computation using the QR decomposition (编) Rao, C.R. Computational Statistics. Handbook of Statistics **9**. North-Holland. 1993. ISBN 0-444-88096-8
- National Physical Laboratory. Chapter 1: Linear Equations and Matrices: Direct Methods. Modern Computing Methods. Notes on Applied Science **16** 2nd. Her Majesty's Stationery Office. 1961

参见

- 方差分析
- 安斯库姆四重奏
- 横截面回归
- 曲线拟合
- 经验贝叶斯方法
- 逻辑斯蒂回归
- M估计
- 非线性回归
- 非参数回归
- 多元自适应回归样条
- Lack-of-fit sum of squares
- 截断回归模型
- 删失回归模型
- 简单线性回归
- 分段线性回归

取自“https://zh.wikipedia.org/w/index.php?title線性回歸&oldid=47275947”