基本回归算法综述

尹达恒
（江南大学物联网工程学院，江苏　无锡）

摘要： 回归分析 (regression analysis) 是确定相互依赖的变量间定量关系的一种统计分析方法。回归算法 (regression algorithm) 是各种回归分析方法在计算机中的具体实现。回归算法是机器学习领域的重要基石，也是当前机器学习界的研究热点之一。本文介绍了回归分析的历史和基本原理，综述了一些主流回归算法的理论和技术，分析了回归算法的研究现状，最后探讨了回归算法今后的研究方向。

关键词：回归分析，回归算法，机器学习

A review on Basic Regression Algorithm

YING Da-heng
(School of IoT, Jiangnan University, Wuxi Jiangsu, China)

Abstract: Regression analysis is a statistical method for estimating the quantitative relationship among interdependent variables, and regression algorithm is a series of implementations of those regression analysis methods in computers. Regression algorithm is an important cornerstone of the machine learning, and is a hot research field in machine learning. This paper introduces the history and the fundamental of regression analysis, reviews the theories and technologies of regression algorithms, analyzed the research state of regression algorithms, and finally discusses the research and development prospects of regression algorithms.

Keywords: regression analysis, regression algorithm, machine learning

1 简介

回归算法 (regression algorithm) 是建立在回归分析 (regression analysis) 理论之上的一系列数据分析算法的总称。一个特定的回归算法代表了一系列对特定自变量和因变量之间的定量关系进行建模、估计和预测的过程。回归算法作为数据分析的一项关键技术，近年来成为人工智能和数据挖掘领域受到普遍重视、研究十分活跃的课题之一。

回归分析最早来源于最小二乘法，分别由法国数学家 Legendre 和德国数学家高斯 Gauss 于 1805 年 [1] 和 1809 年 [2] 独立出版，最初应用于从天文观测数据中确定彗星和小行星轨道的问题。“回归 (regression)”一词的使用可以追溯到十九世纪英国科学家 Francis Galton 对亲子间身高关系的研究，高尔顿用这个词描述父母的身高虽然会遗传给子女，但子女的身高会逐渐“回归到平均身高 (regression to mean)”的现象 [3]。二十世纪初，英国统计学家 Udny Yule 和

Karl Pearson 在高尔顿的研究基础上将“回归”推广为更加一般的统计学概念，将其表述为研究联合分布为高斯分布的反应变量 (response variables) 与解释变量 (explanatory variables) 间关系的方法 [4] [5]。在之后的 1922 年，英国统计学家 Ronald Fisher 提出在回归分析中反应变量与解释变量只需满足边缘分布为高斯分布的条件，弱化了回归分析的限制条件，扩大了回归分析的应用范围 [6]。到二十世纪中叶，随着电子计算机的发展，各种回归分析方法被转化为各种各样的算法并在计算机中得到实现。

时至今日，回归分析仍然是一个热门的研究领域。近半个世纪以来，随着贝叶斯回归理论 [7] [8]、多层次线性模型 [9] 和广义线性模型 [10] 的提出，许多新的回归分析方法应运而生。回归算法的应用领域也逐渐从静态数据分析逐步扩展到时间序列分析 (time series analysis)、图像处理、数据修复等领域。

本文将首先介绍回归分析的基本思想，随后从模型和理论入手介绍一些主流的回归分析方法，并在此基础上对当今广泛使用的基本回归算法进行综述，最后简要探讨回归分析方法未来的发展方向。

2 回归分析的基本思想

回归分析是指已知一个关于自变量 $\boldsymbol{x} \in X$ 和因变量 $y \in Y$ 的满射 $F : X \to Y, y = F(\boldsymbol{x})$ 以及函数 $\hat{y} = f(\boldsymbol{x}, \boldsymbol{\beta})$，求未知变量 $\boldsymbol{\beta}$，使得 $y = F(\boldsymbol{x})$ 与函数 $\hat{y} = f(\boldsymbol{x}, \boldsymbol{\beta})$ 之间的误差 $E = E(F(\boldsymbol{x}), f(\boldsymbol{x}, \boldsymbol{\beta}))$ 最小。

在实际操作中，自变量 $\boldsymbol{x}$ 和未知变量 $\boldsymbol{\beta}$ 可以是数值或向量；求模型误差的方法 $E(,)$ 随实际需求的不同有多种形式；关于自变量 $\boldsymbol{x}$ 和因变量 y 的满射 $F : X \to Y$ 通常由一个样本集合（又称数据集 (data set)）$\mathcal{D} = \{(\boldsymbol{x}, y) | \boldsymbol{x} \in X, y \in Y, y = F(\boldsymbol{x})\}$ 来表示。一般来说，一个能用回归模型有效描述的样本集合应该具有以下特点：

(1) 样本互相独立；
(2) 各自变量间取值互相独立；
(3) 样本分布能代表总体的分布；
(4) 自变量无测量误差；
(5) 因变量测量误差服从正态分布 $\mathcal{N}(0, \sigma^2)$。

当样本集合不能满足其中的某个或某些条件时，进行回归分析就需要使用一些特殊的回归方法，如岭回归和多层次模型回归等。

从本质上讲，回归分析的过程是给定数据集 $\mathcal{D}$ 和函数形式 $f(,)$，通过调节参数 $\boldsymbol{\beta}$，使得在相同的输入 $\boldsymbol{x}$ 下，函数的输出 $\hat{y} = f(\boldsymbol{x}, \boldsymbol{\beta})$ 尽可能地与数据集样

本给出的输出 $y = F(\boldsymbol{x})$ 接近，进而能使用函数的输出 $\hat{y}$ 替代数据集中的 y，从而能通过 $\boldsymbol{\beta}$ 值判断 y 与 $\boldsymbol{x}$ 间的定量关系并对数据集中未出现的数据进行预测。总的来说，回归分析主要解决以下几个问题：

(1) 确定 y 与 $\boldsymbol{x}$ 间的定量关系表达式（回归方程）$\hat{y} = f(\boldsymbol{x}, \boldsymbol{\beta})$；

(2) 对求得的回归方程的可信度（误差）$E(F(\boldsymbol{x}), f(\boldsymbol{x}, \boldsymbol{\beta}))$ 进行评估；

(3) 通过求得的未知参数 β 的值判断 y 与 $\boldsymbol{x}$ 间的定量关系；

(4) 利用求得的回归方程 $\hat{y} = f(\boldsymbol{x}, \boldsymbol{\beta})$ 对数据集中未包含 $(\boldsymbol{x}, y)$ 进行预测。

3 线性回归

在统计学领域中，线性回归是指当函数 $\hat{y} = f(\boldsymbol{x}, \boldsymbol{\beta})$ 关于 $\boldsymbol{x}$ 和 $\boldsymbol{\beta}$ 都是线性函数时进行的回归分析，此时函数 $\hat{y} = f(\boldsymbol{x}, \boldsymbol{\beta})$ 称为一个关于自变量 $\boldsymbol{x}$ 和因变量 y 的线性模型。线性回归中，对未知参数 $\boldsymbol{\beta}$ 的估计通常使用最小二乘法和最大似然估计法。线性回归模型是最简单的回归分析模型，因为简洁直观的特点而成为目前研究最多、使用范围最广的回归分析模型。

3.1 简单线性回归

当自变量向量 $\boldsymbol{x}$ 维数为 1 时，在这样的数据集上进行的线性回归分析称为简单线性回归 (simple linear regression) 或一元线性回归。

简单线性回归的回归方程定义为：

$$\hat{y} = f(\boldsymbol{x}, \boldsymbol{\beta}) = \beta_0 + x\beta_1 \tag{1}$$

简单线性回归使用普通最小二乘法 (ordinary least squares ,OLS) 进行参数估计。在使用普通最小二乘法求简单线性回归模型的参数过程中，模型的误差函数 $E(,)$ 被定义为残差平方和 (residual sum of squares,SSR)，即：

$$E = \sum_{(\boldsymbol{x},y)\in\mathcal{D}} (f(\boldsymbol{x}, \boldsymbol{\beta}) - y)^2 = \sum_{(x,y)\in\mathcal{D}} (\beta_0 + x\beta_1 - y)^2 \tag{2}$$

为了使模型的残差平方和达到最小，将 E 分别对 β_0 和 β_1 求偏导得：

$$\frac{\partial E}{\partial \beta_0} = 2 \sum_{(x,y)\in\mathcal{D}} (\beta_0 + x\beta_1 - y) \tag{3}$$

$$\frac{\partial E}{\partial \beta_0} = 2 \sum_{(x,y)\in\mathcal{D}} (\beta_0 + x\beta_1 - y)\, x \tag{4}$$

令式 (3) 和式 (4) 为零，可解得使残差平方和 E 达到最小的 β_0 和 β_1：

$$\beta_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} \qquad ((x,y)\in\mathcal{D}) \tag{5}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \qquad ((x,y)\in\mathcal{D}) \tag{6}$$

式 (1) 结合式 (5) 和式 (6) 即简单线性回归模型的表达式。简单线性回归是历史上出现的第一个回归模型，是回归分析的根源所在。

3.2 一般线性回归

一般线性模型 (general linear model) 又称为多元回归模型 (multivariate regression model)，是简单线性模型在向量 $\boldsymbol{x}$ 维数大于 1 时的推广，模型表述为：

$$\hat{y} = f(\boldsymbol{x}, \boldsymbol{\beta}) = \boldsymbol{x}^T\boldsymbol{\beta} = \sum_{i=0}^{n} x_i\beta_i = \beta_0 + \sum_{i=1}^{n} x_i\beta_i \tag{7}$$

其中 β_0 为偏置 (bias) 项，$x_0 = 1$ 为一恒定量。

一般线性回归通常使用普通最小二乘法的矩阵形式进行参数估计。矩阵形式的普通最小二乘法中残差平方和表达式为：

$$E = \sum_{(\boldsymbol{x},y)\in\mathcal{D}} (f(\boldsymbol{x},\boldsymbol{\beta}) - y)^2 = \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2 \tag{8}$$

其中 $\boldsymbol{y}$ 是由样本数据集中的因变量组成的向量、$\boldsymbol{X}$ 是由样本数据集中的因变量组成的矩阵，形式为：

$$\begin{gathered}
\boldsymbol{y} = (y_1, y_2, \ldots y_i, \ldots y_m)^T \\
\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x_1^T} & 1 \\ \boldsymbol{x_2^T} & 1 \\ \cdots & \cdots \\ \boldsymbol{x_i^T} & 1 \\ \cdots & \cdots \\ \boldsymbol{x_m^T} & 1 \end{pmatrix} \\
(\boldsymbol{x_i}, y_i) \in \mathcal{D}
\end{gathered} \tag{9}$$

为了使模型的残差平方和最小，将 E 对 $\boldsymbol{\beta}$ 求偏导可得：

$$\frac{\partial E}{\partial \boldsymbol{\beta}} = 2\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}) \tag{10}$$

当 $\boldsymbol{X}^T\boldsymbol{X}$ 为满秩矩阵时，令式 (10) 为 0 可解得：

$$\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{11}$$

式 (7) 结合式 (11) 即一般线性回归模型的表达式 [11]。

除最小二乘法之外，一般线性回归的参数估计还有一种最大似然法 (maximum likelihood,ML)。最大似然法的思想是在给定数据集合 $\mathcal{D}$ 的条件下调节参数 $\boldsymbol{\beta}$，使数据集合 $\mathcal{D}$ 出现的概率最大。首先，可以计算出在给定参数 $\boldsymbol{\beta}$ 时数据集合 $\mathcal{D}$ 出现的概率：

$$L(\boldsymbol{\beta}) = p(\mathcal{D}|\boldsymbol{\beta}) = \prod_{(\boldsymbol{x},y)\in\mathcal{D}} p((\boldsymbol{x},y)|\boldsymbol{\beta}) \tag{12}$$

概率 $L(\boldsymbol{\beta})$ 又称为似然函数，是最大似然估计要最大化的目标。为方便计算，一般将似然函数取对数再求偏导数零点：

$$\frac{\partial ln(L(\boldsymbol{\beta}))}{\partial \boldsymbol{\beta}} = \sum_{(\boldsymbol{x},y)\in\mathcal{D}} \frac{\partial ln(p((\boldsymbol{x},y)|\boldsymbol{\beta}))}{\partial \boldsymbol{\beta}} = 0 \tag{13}$$

解方程 (13) 即可求出最大似然估计的参数 $\boldsymbol{\beta}$。

3.3 贝叶斯线性回归

贝叶斯线性回归 (Bayesian linear regression) [7] [8] 是指使用贝叶斯推断 (Bayesian inference) 求解的一般线性回归。从贝叶斯学派的观点来看，构建线性回归应使用概率分布而非点估计方法，其一般模型表达式如下：

$$\hat{y} \sim \mathcal{N}(\boldsymbol{x}^T\boldsymbol{\beta}, \sigma^2\boldsymbol{I}) \tag{14}$$

可以看出，与一般线性回归的模型表达式 (7) 相比，贝叶斯线性回归的因变量预测值 $\hat{y}$ 不再表示单个数值，而是表示一个均值为 $\boldsymbol{x}^T\boldsymbol{\beta}$、方差为 $\sigma^2\boldsymbol{I}$ 的正态分布。

按照连续概率分布的积分形式写出贝叶斯回归模型的表达式为：

$$\hat{y} \sim p(\hat{y}|\boldsymbol{x},\mathcal{D}) = \int_{\boldsymbol{\beta}} p(\hat{y}|\boldsymbol{x},\boldsymbol{\beta})p(\boldsymbol{\beta}|\mathcal{D})d\boldsymbol{\beta} \tag{15}$$

贝叶斯线性回归的参数估计一般采用最大化 $p(\boldsymbol{\beta}|\mathcal{D})$ 的最大后验概率法 (maxi-

mum a posteri,MAP)。在最大后验概率法中，最大化目标后验概率由下式给出：

$$p(\boldsymbol{\beta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\beta})p(\boldsymbol{\beta})}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\boldsymbol{\beta})p(\boldsymbol{\beta})}{\int_{\boldsymbol{\beta}} p(\mathcal{D}|\boldsymbol{\beta})p(\boldsymbol{\beta})d\boldsymbol{\beta}} \tag{16}$$

其中 $p(\boldsymbol{\beta})$ 是一个人为指定的关于参数 $\boldsymbol{\beta}$ 的一个先验概率，这个概率值表征了人们对参数 $\boldsymbol{\beta}$ 的已有知识和经验。与 3.2中计算最大似然估计的方法类似，先对后验概率取对数：

$$\begin{aligned} ln(p(\boldsymbol{\beta}|\mathcal{D})) &= ln(p(\mathcal{D}|\boldsymbol{\beta})) + ln(p(\boldsymbol{\beta})) - ln(p(\mathcal{D})) \\ &= ln\left(\prod_{(\boldsymbol{x},y)\in\mathcal{D}} p((\boldsymbol{x},y)|\boldsymbol{\beta})\right) + ln(p(\boldsymbol{\beta})) - ln(p(\mathcal{D})) \end{aligned} \tag{17}$$

与方程 (13) 类似，进一步将式 (17) 对 $\boldsymbol{\beta}$ 求偏导零点, 可以得到方程：

$$\frac{\partial ln(p(\boldsymbol{\beta}|\mathcal{D}))}{\partial \boldsymbol{\beta}} = \sum_{(\boldsymbol{x},y)\in\mathcal{D}} \frac{\partial ln(p((\boldsymbol{x},y)|\boldsymbol{\beta}))}{\partial \boldsymbol{\beta}} + \frac{\partial ln(p(\boldsymbol{\beta}))}{\partial \boldsymbol{\beta}} = 0 \tag{18}$$

求解方程 (18) 即可得到最大后验概率法的估计结果。

对比式 (13) 可以发现，相比于最大似然估计的方法，贝叶斯方法的最大后验参数估计方程中多了一项先验概率偏导 $\frac{\partial ln(p(\boldsymbol{\beta}))}{\partial \boldsymbol{\beta}}$，表明最大后验参数估计不仅从数据集合本身提取信息，还将人们的已有经验纳入考虑，因此在一些场合下有比最大似然估计方法更高的准确率。但是在实际应用中，先验概率 $p(\boldsymbol{\beta})$ 有时并不能很好的总结出来，因此除最大后验参数法外，贝叶斯线性回归还会采用一种称为贝叶斯增量学习 [12] 的方法，将前一步的后验概率作为下一步的先验概率，逐步进行回归。假设样本集合 $\mathcal{D}$ 中有 n 个样本，则当 $n > 1$ 时有：

$$p(\mathcal{D}_i|\boldsymbol{\beta}) = p((\boldsymbol{x}_i, y_i)|\boldsymbol{\beta})p(\mathcal{D}_{i-1}|\boldsymbol{\beta}) \qquad i \in \{1, \ldots, n\} \tag{19}$$

将式 (19) 带入式 (16) 可得：

$$\begin{aligned} p(\boldsymbol{\beta}|\mathcal{D}_i) &= \frac{p(x_i|\boldsymbol{\beta})p(\mathcal{D}_{i-1}|\boldsymbol{\beta})p(\boldsymbol{\beta})}{\int_{\boldsymbol{\beta}} p(x_i|\boldsymbol{\beta})p(\mathcal{D}_{i-1}|\boldsymbol{\beta})p(\boldsymbol{\beta})d\boldsymbol{\beta}} \\ &= \frac{p(x_i|\boldsymbol{\beta})p(\boldsymbol{\beta}|\mathcal{D}_{i-1})}{\int_{\boldsymbol{\beta}} p(x_i|\boldsymbol{\beta})p(\boldsymbol{\beta}|\mathcal{D}_{i-1})d\boldsymbol{\beta}} \end{aligned} \tag{20}$$

相当于 $p(\boldsymbol{\beta}|\mathcal{D}_{i-1})$ 成为了 $p(\boldsymbol{\beta}|\mathcal{D}_i)$ 的先验概率分布，即：

$$p(\boldsymbol{\beta}|\mathcal{D}_i) \propto p(x_n|\theta)p(\boldsymbol{\beta}|\mathcal{D}_{i-1}) \tag{21}$$

贝叶斯增量学习对数据有自适应能力，可以在防止过拟合的情况下重复利用实验数据，并且增量学习的模式和机器学习中在线学习的思想高度重合，因此在机器学习领域广受推崇。

3.4 岭回归

岭回归 (ridge regression) 又称吉洪诺夫正则化 (Tikhonov regularization) 在机器学习领域中又称权重衰减 (weight decay)，是 Andrey Tikhonov 于二十世纪四十年代发明并推广的针对不适定问题 (ill-posed problem) 进行回归分析的一种正则化方法 [13]。在一般线性模型中，如果式 (10) 中的 $\boldsymbol{X}^T\boldsymbol{X}$ 不为满秩矩阵或自变量之间具有较强的线性相关性时，这个一般线性模型就成为一个不适定问题，此时 $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ 误差很大或无法计算，传统的最小二乘法缺乏稳定性与可靠性。为了将不适定问题转化为适定问题，Tikhonov 提出可以为均方误差 E 加上一个正则化 (regularization) 项，变为：

$$E = \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2 + \|\boldsymbol{\Gamma}\boldsymbol{\beta}\|^2 \tag{22}$$

运用式 (22) 和 3.2中求偏导的方法可以解出岭回归模型的估计参数表达式：

$$\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Gamma})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{23}$$

其中 $\boldsymbol{\Gamma}$ 称为 Tikhonov 矩阵，该矩阵是根据实际情况人为选定的。在大多数情况下，岭回归选择 $\boldsymbol{\Gamma} = \lambda\boldsymbol{I}$，其中 λ 为一人为选择的常数，$\boldsymbol{I}$ 为一单位矩阵，此时的正则化项变为 $\|\boldsymbol{\Gamma}\boldsymbol{\beta}\|^2 = \lambda\|\boldsymbol{\beta}\|^2$，这样的正则化项称为 L_2 范数 (L_2 norm)，这种形式的岭回归也被称为 L_2 范数正则化 (L_2 regularization)，在机器学习领域具有广泛应用 [14]。岭回归是对最小二乘回归的一种补充，它舍弃了最小二乘法求解线性回归的无偏性来换取较高的稳定性，能适用于很多最小二乘回归无法完成或是精度很低的问题，且能保证较高的计算精度 [15]。

4 分层线性模型

分层线性模型 (multilevel linear models) [16] [9] 又称等级线性模型 (hierarchical linear models,HLM) [17]、随机系数模型 (random coefficient models) [18]、协方差成分模型 (covariance components models) [19], 是针对经典线性回归分析在处理具有多层结构的数据时所存在的不足而提出的。在科学研究中的很多问题都表现为多水平的、多层次的数据结构, 这些分层结构数据主要可以分为六大类 [20]：

(1) 空间分层数据 (hierarchical data)；

(2) 时间纵向数据 (longitudinal data)；

(3) 重复测量数据 (repeated measurement data)；

(4) 广义聚类数据 (generalized clustered data)；

(5) 名义分类数据 (norminal categorical data)；

(6) 有序分类数据 (ordinal categorical data)。

其中一个最为典型的例子就是教育研究中学生、班级、学校互相嵌套的空间分层结构。在这种关系下, 模型中除了要包含学生的个体情况外，不同班级学生的某些变量会受到班级的影响，同一个班级的学生又有某些变量取值相近，因此从总体上对学生情况的取样分析无法满足回归分析中样本取值互相独立的假设；而若是将学生的特征集中到班级上从而对班级进行分析，又会由于忽略班级中学生自身情况的不同而导致统计结果存在一定偏差 [21]，因此，单一层次的回归分析不能很好地解决多层次嵌套的问题。

为了单一层次回归分析存在的问题，多层次线性模型将个体（学生）分析和组（班级）分析区别开来，将样本按照实际情况分为组内分析和组间分析两个分析层次。假设有 $(\boldsymbol{x}, \boldsymbol{w}, y)$ 的独立同分布的观察数据集合 $\mathcal{D} = \{(\boldsymbol{x}_{i,j}, \boldsymbol{w}_j, y_{i,j})\}$，其中 i 为个体标号，j 为组标号，$\boldsymbol{x}_{i,j}$ 为个体的特征向量，$\boldsymbol{w}_j$ 为组的特征向量，y_i 为因变量。多层次模型中第一个层次组内分析的模型表达式为 [20]：

$$\hat{y}_{i,j} = \boldsymbol{\beta}_j \boldsymbol{x}_{i,j} + \varepsilon_i = \beta_{0,j} + \sum_k \beta_{k,j} x_{k,i,j} + \varepsilon_i \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{24}$$

式中 k 为个体特征标号，$\beta_{k,j}$ 表征第 k 个个体特征在第 j 组的水平预测变量对于因变量的效应，ε_i 是不可观测的随机效应 (random effects) 变量，假定其与 $\boldsymbol{x}_{i,j}$ 独立且服从方差为 0 的正态分布。

组间分析是多层次分析模型中的第二层次，模型表达式为 [20]：

$$\boldsymbol{\beta}_j = \boldsymbol{\Gamma} \boldsymbol{w}_j + \boldsymbol{u}_j \qquad \boldsymbol{u}_j \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{T}) \tag{25}$$

其中 $\boldsymbol{\Gamma}$ 矩阵表征系统的固定效应 (fixed effects)，$\boldsymbol{u}_j$ 为第二层次的随机效应变量，假定其与 $\boldsymbol{w}_j$ 独立且服从均值向量为 $\boldsymbol{0}$，协方差矩阵为 T 的多元分布 [20]。方程 (25) 中的 $\boldsymbol{\beta}_j$ 由方程 (24) 使用最小二乘 (3.2) 等参数估计方法估计得出 [17]；对于方程 (25) 的回归实际上是一个多因变量的回归问题，可以讲其分解为多个单因变量回归问题求解。

多层次模型的优点主要有三个：首先它能通过在其他组中存在的相似估计对单个组提出更好的估计；其次是能显示出组级别的因素对个体的影响；再者它

可以分离各分析水平内的方差成分 [21]。多层次模型因为其天生的对复杂嵌套问题的适应性而成为社会科学和系统生物学等领域的常用模型之一。

5 广义线性模型

线性回归模型主要适用于因变量为连续型随机变量且与自变量成线性关系的情况，而 Nelder 等人于 1972 年 [10] 提出的广义线性模型 (generalized linear models,GLM) 通过指定一个连接函数将因变量的数学期望与自变量的线性组合联系起来，并且将因变量的分布推广到广义指数族分布，能在不强行改变数据的自然度量的情况下使回归分析结果可以具有非线性和非恒定方差结构 [22]。广义线性模型从以下两个方面推广了 3.2中所展示的一般线性回归模型：

(1) 数据集 $\mathcal{D}$ 中样本因变量 y 的分布由正态分布推广到指数族分布:

$$y \sim f(y|\theta,\phi) = e^{\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)} \tag{26}$$

其中 $a(\cdot)$，$b(\cdot)$，$c(,)$ 为已知的连续函数，θ 和 ϕ 为未知参数。

(2) 自变量 $\boldsymbol{x}$ 和因变量 y 之间的关系由线性函数推广到一个单调可微函数。

设 $(\boldsymbol{x},y) \in \mathcal{D}$ 为样本集合中的一个观测值，因变量 y 满足一个指数族分布，则广义线性模型的一般表达式如下：

$$\begin{aligned} \eta &= g(\mu) = \boldsymbol{x}^T\boldsymbol{\beta} \\ \mu &= E(y) \end{aligned} \tag{27}$$

其中 $g(\cdot)$ 表示一个单调可微函数，称为连接函数 (link function)。根据连接函数的不同，广义线性回归模型具有不同的形式。广义线性回归中所使用的连接函数一般为典则连接函数 (canonical link function) [22]，即连接函数 $g(\cdot)$ 满足：

$$g(E(y)) = g(\mu) = \theta \tag{28}$$

其中 $y \sim f(y|\theta,\phi) = e^{\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)}$。典型的典则连接函数形式如表 1所示 [22] [23]。可以看出，在广义线性模型下，经典线性模型是在 $y \sim \mathcal{N}(\mu,\sigma^2)$ 时的特例。

广义线性回归的参数估计一般采用似然方程的迭代加权最小二乘解法。在数据集合 $\mathcal{D}$ 上，未知参量 $\boldsymbol{\beta}$ 的对数似然函数为：

$$lnL(\boldsymbol{\beta}) = ln\left(\prod_{(\boldsymbol{x},y)\in\mathcal{D}} f(y|\theta,\phi)\right) = \sum_{(\boldsymbol{x},y)\in\mathcal{D}}\left(\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)\right) \tag{29}$$

表 1: 典型典则连接函数

y 的分布	连接函数	回归方程形式
正态分布 $y \sim \mathcal{N}(\mu, \sigma^2)$	单位函数 $g(\mu) = \mu$	$\mu = \boldsymbol{x}^T\boldsymbol{\beta}$
指数分布 $y \sim E(\mu)$ Gamma 分布 $y \sim \Gamma(\alpha, \mu)/\alpha$	负倒数 $g(\mu) = \frac{1}{\mu}$	$\mu = (\boldsymbol{x}^T\boldsymbol{\beta})^{-1}$
逆高斯分布 $y \sim IG(\mu, \lambda)$	负倒数 $g(\mu) = \frac{1}{\mu^2}$	$\mu = (\boldsymbol{x}^T\boldsymbol{\beta})^{-\frac{1}{2}}$
泊松分布 $y \sim P(\mu)$	对数函数 $g(\mu) = ln(\mu)$	$\mu = e^{(\boldsymbol{x}^T\boldsymbol{\beta})}$
二项分布 $y \sim B(m, \mu)/m$	Logistic 函数 $g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{1}{1+e^{-\boldsymbol{x}^T\boldsymbol{\beta}}}$

令偏导数为 0，即：

$$\frac{\partial lnL(\boldsymbol{\beta})}{\partial \beta_r} = \frac{\partial}{\partial \beta_r} \sum_{(\boldsymbol{x}.y)\in\mathcal{D}} \left(\frac{y\theta - b(\theta)}{a(\phi)}\right) \tag{30}$$

由典则连接函数的性质可以推出 [22]：

$$\sum_{(\boldsymbol{x}.y)\in\mathcal{D}} \frac{(y-\mu)x_i}{aV(\mu)g'(\mu)} = 0 \tag{31}$$

其中 $V(\mu) = b''(\theta)$。在一般情况下，似然方程 (31) 是关于 $\boldsymbol{\beta}$ 的非线性方程组，无法给出显式解，只能通过迭代方法求解。迭代加权最小二乘解法的迭代方程式为 [22]：

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{(m+1)} &= \left(\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{Z}^{(m)} \\
\boldsymbol{W}^{(m)} &= Diag(\omega_1^{(m)}, \ldots, \omega_n^{(m)}) \\
\omega_i^{(m)} &= \frac{1}{a_i V(\mu_i^{(m)}) g'(\mu_i^{(m)})^2} \\
\boldsymbol{Z}^{(m)} &= (Z_1^{(m)}, \ldots, Z_n^{(m)})^T \\
Z_i^{(m)} &= g(\mu_i^{(m)}) + \left(y_i - \mu_i^{(m)}\right) g'\left(\mu_i^{(m)}\right) \\
\boldsymbol{\mu}^{(m)} &= (\mu_1^{(m)}, \ldots, \mu_n^{(m)})^T = g^{-1}(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(m)})
\end{aligned} \tag{32}$$

其中数据集合 $\mathcal{D}$ 表述为 $\mathcal{D} = \{(\boldsymbol{x}_i, y_i) | 1 \le i \le n\}$。迭代过程的初值为 [22]：

$$\boldsymbol{\mu}^{(0)} = (\mu_1^{(0)}, \ldots, \mu_n^{(0)})^T = \boldsymbol{y} = (y_1^{(0)}, \ldots, y_n^{(0)})^T \tag{33}$$

广义线性模型可以应用于众多领域的数据分析问题，除作为特例的经典线性模型外，广义线性模型中 $y \sim P(\mu)$ 的泊松分布模型和 $y \sim B(m, \mu)/m$ 的二项分布模型（又称逻辑回归 (logistic regression)）在医学、社会学、经济学以及人工智能领域已经得到广泛应用。

6 总结

在简单线性回归提出至今的二百多年时间里，对于回归分析方法和回归算法的研究获得了长足的发展。各类相关信息的多样化、多层次、多角度回归分析研究以及突破一般线性回归的假设开发新的回归分析方法，是近代回归分析研究的突出发展方向。能否有效综合样本信息构建恰当的回归模型、能否从低质甚至混乱的数据中提取出正确的信息，是决定回归分析效果的核心问题。近年来，各类新兴建模技术的发展有效促进了回归分析技术的提升，深度学习是其中的典型代表。逻辑回归、岭回归、分层线性回归等回归分析算法和梯度下降 (Gradient Descent) 及其衍生出的各类最优化算法在深度学习领域的成功结合曾一度让深度学习领域取得飞跃性的研究进展，使人工智能在越来越多的场景中表现出超越人类的能力，也使得与深度学习相关的回归算法和优化技术成为目前最受关注的发展方向。

本文针对目前常见的回归分析算法的基本思想方法进行了综述研究，将所有方法按照核心技术差异进行了归类介绍，并适当描述了各类方法之间、同类方法之间的一些差别与联系。本文还分析统计了各类方法的部分理论细节，并在最后分析了回归分析的主要发展趋势。回归分析属于基础研究，是对科技发展有重要推动作用的基础科学的一部分，具有重大的实用价值，针对具体应用领域的研究则更加重要, 以便促进我国高新技术的迅速发展。希望本文能对相关科研人员了解回归分析和回归算法并开展相关研究起到微薄的作用。

参考文献

[1] A.M. Legendre. Nouvelles méthodes pour la détermination des orbites des comètes. *Sur la Méthode des moindres quarrés*, 1805.

[2] C.F. Gauss. *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientum.* 1809.

[3] Francis Galton. Typical laws of heredity. *Nature*, 15:492–495, 512–514, 532–533, 1877.

[4] G. Udny Yule. On the theory of correlation. *Journal of the Royal Statistical Society*, 60(4):812–854, 1897.

[5] KARL PEARSON. The law of ancestral heredity*. *Biometrika*, 2(2):211–228, 1903.

[6] R. A. Fisher. The goodness of fit of regression formulae, and the distribution of regression coefficients. *Journal of the Royal Statistical Society*, 85(4):597–612, 1922.

[7] A. F. M. Smith. A general bayesian linear model. *Journal of the Royal Statistical Society*, 35(1):67–75, 1973.

[8] A. F. M. Lindley, D. V.，Smith. Bayes estimates for the linear model. *Journal of the Royal Statistical Society*, 34(1):1–41, 1972.

[9] George Y，Entwisle Barbara Mason, William M，Wong. Contextual analysis through the multilevel linear model. *Sociological methodology*, 14:72–103, 1983.

[10] Robert WM Nelder, John Ashworth，Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

[11] 周志华. 机器学习. 清华大学出版社, 2016.

[12] Rob，Perona Pietro Fei-Fei, Li，Fergus. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer vision and Image understanding*, 106(1):59–70, 2007.

[13] Andrei Nikolaevich Tikhonov. On the solution of ill-posed problems and the method of regularization. In *Doklady Akademii Nauk*, volume 151, pages 501–504. Russian Academy of Sciences.

[14] Andrew Y Ng. Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, page 78. ACM.

[15] 郭鹏妮. 岭回归与分位数回归的研究及结合应用. 硕士, 2014.

[16] Harvey Goldstein. *Multilevel statistical models*, volume 922. John Wiley & Sons, 2011.

[17] Anthony S Raudenbush, Stephen W，Bryk. *Hierarchical linear models: Applications and data analysis methods*, volume 1. Sage, 2002.

[18] Nicholas T Longford. Logistic regression with random coefficients. *Computational Statistics & Data Analysis*, 17(1):1–15, 1994.

[19] Donald B，Tsutakawa Robert K Dempster, Arthur P，Rubin. Estimation in covariance components models. *Journal of the American Statistical Association*, 76(374):341–353, 1981.

[20] 田茂再. 高等分层分位回归建模理论. 科学出版社, 2015.

[21] 雷雳，张雷. 多层线性模型的原理及应用. 首都师范大学学报 *(*社会科学版*)*, (02):110–114, 2002.

[22] 梅长林，王宁. 近代回归分析方法. 北京*:* 科学出版社, 2012.

[23] 胡宏昌，崔恒建，秦永松等. 近代线性回归分析方法, 2012.