Ch13: Linear and Logistic Regression–多元回归

王 成

chengwang@sjtu.edu.cn

上海交通大学数学科学学院

自我介绍

王 成: 数学科学学院,
Email: chengwang@sjtu.edu.cn， Office: 理化大楼6-525;

自我介绍

王 成: 数学科学学院,
Email: chengwang@sjtu.edu.cn， Office: 理化大楼6-525;

主要经历:

- 2003.9-2013.7, 中国科学技术大学, 本、硕、博;
- 2013.9-2014.8, 香港浸会大学, 博士后研究员;
- 2014.9-2021.12, 上海交通大学, 长聘教规副教授;
- 2022.1-至今, 上海交通大学, 长聘副教授。

主要研究方向: 随机矩阵、高维数据的统计推断、统计优化算法。

Section 1

多元回归模型

回归样本

给定样本

$$(\mathbf{X}_1, Y_1), \quad \cdots, (\mathbf{X}_n, Y_n),$$

其中$\mathbf{X}_1, \cdots, \mathbf{X}_n \in \mathbb{R}^p$为解释变量，$Y_1, \cdots, Y_n \in \mathbb{R}$ 为响应变量.

最小二乘估计

把样本$(\mathbf{X}_i, Y_i)$看成$\mathbb{R}^{p+1}$空间中点，考虑最小二乘估计

$$\arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i^\top \beta)^2. \tag{1}$$

最小二乘估计

记

$$\mathbb{X} = \begin{pmatrix} \mathbf{X}_1^\top \\ \vdots \\ \mathbf{X}_n^\top \end{pmatrix}_{n\times p} = \begin{pmatrix} X_{11} & \dots & X_{1p} \\ \vdots & \vdots & \vdots \\ X_{n1} & \dots & X_{np} \end{pmatrix}, \mathbb{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix},$$

用矩阵形式优化问题(1)可以表示成

$$\arg\min_{\beta\in\mathbb{R}^p}(\mathbb{Y}-\mathbb{X}\beta)^\top(\mathbb{Y}-\mathbb{X}\beta) = \arg\min_{\beta\in\mathbb{R}^p}\frac{1}{2n}(\mathbb{Y}-\mathbb{X}\beta)^\top(\mathbb{Y}-\mathbb{X}\beta).$$

注意

实际使用中，设定数据矩阵的第一列即$X_{11},\dots,X_{n1}$为1，这样可以自动的把常数项包括进模型.

最小二乘估计

求导数，

$$\frac{\partial(\mathbb{Y}-\mathbb{X}\beta)^\top(\mathbb{Y}-\mathbb{X}\beta)}{\partial\beta} = 2\mathbb{X}^\top(\mathbb{X}\beta-\mathbb{Y}),$$
$$\frac{\partial^2(\mathbb{Y}-\mathbb{X}\beta)^\top(\mathbb{Y}-\mathbb{X}\beta)}{\partial\beta\partial\beta^\top} = 2\mathbb{X}^\top\mathbb{X} \geq 0.$$

当矩阵$\mathbb{X}^\top\mathbb{X}$ 严格正定的时候，我们有最小二乘估计

$$\hat{\beta} = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\mathbb{Y}.$$

最小二乘估计

求导数，

$$\frac{\partial(\mathbb{Y}-\mathbb{X}\beta)^\top(\mathbb{Y}-\mathbb{X}\beta)}{\partial\beta} = 2\mathbb{X}^\top(\mathbb{X}\beta-\mathbb{Y}),$$

$$\frac{\partial^2(\mathbb{Y}-\mathbb{X}\beta)^\top(\mathbb{Y}-\mathbb{X}\beta)}{\partial\beta\partial\beta^\top} = 2\mathbb{X}^\top\mathbb{X} \geq 0.$$

当矩阵$\mathbb{X}^\top\mathbb{X}$ 严格正定的时候，我们有最小二乘估计

$$\hat{\beta} = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\mathbb{Y}.$$

思考

$\mathbb{X}^\top\mathbb{X}$不严格正定会发生什么情况?

Section 2

应用：Galton数据集

Galton数据集

Francis Galton(1886)调查了英国205个家庭，得到928个成年孩子的身高数据. 具体数据格式为:

- Family: The family that the child belongs to, labeled from 1 to 204 and 136A
- Father: The father's height, in inches
- Mother: The mother's height, in inches
- Gender: The gender of the child, male (M) or female (F)
- Height: The height of the child, in inches
- Kids: The number of kids in the family of the child

R软件中可以通过R Package "HistData" 直接调用Galton的数据集.

Galton数据集

```
rm(list=ls()) ###初始化编程环境
set.seed(123) ##设置随机种子
library(HistData)
data(GaltonFamilies)
galton0=GaltonFamilies
summary(galton0)

##      family          father          mother       midparentHeight     children
##  185     : 15   Min.   :62.0   Min.   :58.00   Min.   :64.40   Min.   : 1.000
##  066     : 11   1st Qu.:68.0   1st Qu.:63.00   1st Qu.:68.14   1st Qu.: 4.000
##  120     : 11   Median :69.0   Median :64.00   Median :69.25   Median : 6.000
##  130     : 11   Mean   :69.2   Mean   :64.09   Mean   :69.21   Mean   : 6.171
##  166     : 11   3rd Qu.:71.0   3rd Qu.:65.88   3rd Qu.:70.14   3rd Qu.: 8.000
##  097     : 10   Max.   :78.5   Max.   :70.50   Max.   :75.43   Max.   :15.000
##  (Other):865
##     childNum          gender      childHeight
##  Min.   : 1.000   female:453   Min.   :56.00
##  1st Qu.: 2.000   male  :481   1st Qu.:64.00
##  Median : 3.000                Median :66.50
##  Mean   : 3.586                Mean   :66.75
##  3rd Qu.: 5.000                3rd Qu.:69.70
##  Max.   :15.000                Max.   :79.00
##
```

Galton数据集-预处理

```
galton1=galton0
##原始数据中的英寸转为厘米
galton1[,c(2,3,8)]=2.54*galton0[,c(2,3,8)]
##女性数据乘以1.08消除性别影响
galton1[,3]=1.08*galton1[,3];
galton1[galton1[,7]==c("female"),8]=
  1.08*galton1[galton1[,7]==c("female"),8]
##保持整理好的数据集
write.csv(galton1,file='GaltonHeight.csv')
```

Galton数据集-lm建模

```
## 解释变量：父母平均身高(母亲数据乘以1.08)
X<-(galton1[,2]+galton1[,3])/2
## 响应变量：成年子女身高(女性同样乘以1.08)
Y<-galton1[,8]
##基于R自带lm函数计算线性模型
lmobj<-lm(Y~X)
##展示模型结果
summary(lmobj)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.1279  -3.7989   0.2384   3.9026  23.1909
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 50.59048    7.16609    7.06 3.26e-12 ***
## X            0.71258    0.04075   17.49  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.699 on 932 degrees of freedom
## Multiple R-squared:  0.247,Adjusted R-squared:  0.2462
## F-statistic: 305.8 on 1 and 932 DF,  p-value: < 2.2e-16
```

Galton数据集-最小二乘法

```
n<-dim(galton1)[1]
##构建解释变量矩阵，第一列人为增加一列全是1的的元素
X1<-matrix(1,nrow =n,ncol =2)
X1[,2]=X;
##基于公式计算最小二乘估计
beta1<-solve(t(X1)%*%X1,t(X1)%*%Y)
 ##展示模型结果并与lm结果进行比较
beta1

##              [,1]
## [1,] 50.590482
## [2,]  0.712585
```

Galton数据集-预测

```
## 计算预测值
Y1<-as.numeric(t(beta1)%*%c(1,(226+1.08*190)/2))
Y1

## [1] 204.2238

Y1/1.08

## [1] 189.0961
```

Galton数据集-多元模型

```
##我们以父亲身高、母亲身高作为解释变量，考虑一个多元回归模型
n<-dim(galton1)[1]
X2<-matrix(1,nrow =n,ncol =3) ##构建解释变量矩阵，第一列人为增加一列全是1的
的元素
X2[,2]=galton1[,2];
X2[,3]=galton1[,3];
beta2<-solve(t(X2)%*%X2,t(X2)%*%Y) ##基于公式计算最小二乘估计
beta2                 ##展示模型结果

##             [,1]
## [1,] 50.6093753
## [2,]  0.4087058
## [3,]  0.3037864

lm(Y~galton1[,2]+galton1[,3])

##
## Call:
## lm(formula = Y ~ galton1[, 2] + galton1[, 3])
##
## Coefficients:
##  (Intercept)  galton1[, 2]  galton1[, 3]
##      50.6094        0.4087        0.3038
```

Galton数据集-多元模型预测

```r
##基于多元回归模型进行预测
Y2<-as.numeric(t(beta2)%*%c(1,226,190*1.08))  ## 计算预测值
Y2

## [1] 205.3138

Y2/1.08

## [1] 190.1054
```

Section 3

延伸

回归模型的延伸

- 多项式回归模型：

$$Y = \alpha + \beta_1 X + \cdots + \beta_k X^k + \epsilon;$$

- 变量变换: 例如股票数据

$$Y = \alpha + \beta \log X + \epsilon \quad \text{or} \quad \log Y = \alpha + \beta X + \epsilon,$$

- 局部线性回归和局部多项式回归

$$\arg\min_{\beta} \sum_{j=1}^{n} w_i (Y_i - X_i^\top \beta)^2.$$

Section 4

如何理解线性模型

线性模型的参数估计

线性模型是监督学习中最重要的统计模型，基于线性模型可以延伸得到很多重要的其他模型。

思考

数理统计中参数估计方法学习过矩估计方法和极大似然估计方法，你认为最小二乘法属于那种方法?

参数极大似然估计

在数理统计中，对于来自某一总体的样本

$$X_1, \cdots, X_n \ i.i.d. \ \sim f(x, \theta).$$

对于未知参数θ, 极大似然估计(Maximum Likelihood Estimation, MLE):

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{n} f(X_i, \theta).$$

参数极大似然估计

在极大似然估计中，下式称为似然函数(Likelihood function):

$$L(\theta) = \prod_{i=1}^{n} f(X_i, \theta).$$

注意到log(·)是单调增函数，很多时候也会考虑对数极大似然估计(log-likelihood):

$$\hat{\theta} = \arg\max_{\theta} \log\{\prod_{i=1}^{n} f(X_i, \theta)\} = \arg\max_{\theta} \sum_{i=1}^{n} \log f(X_i, \theta).$$

参数极大似然估计

类似于最小二乘法，从损失函数的角度理解MLE:

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{n} f(X_i, \theta) = \arg\max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log f(X_i, \theta)$$

$$= \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} -\log f(X_i, \theta).$$

可以从似然函数导出合适的损失函数(负对数似然–Negative log-likelihood).

参数极大似然估计

类似于最小二乘法，从损失函数的角度理解MLE:

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{n} f(X_i, \theta) = \arg\max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log f(X_i, \theta)$$
$$= \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} -\log f(X_i, \theta).$$

可以从似然函数导出合适的损失函数(负对数似然–Negative log-likelihood).

注意

极大似然估计是统计学中非常重要的一种参数估计方法，可以导出很多重要实用的统计模型.

Pros and Cons

Pros:

- 极大似然估计理论上是最有效的估计.
- 极大似然估计的构造思想非常简洁,是统计思想的最好体现. 在很多参数估计或者模型估计中，极大似然估计的思想被大量使用.

Pros and Cons

Pros:

- 极大似然估计理论上是最有效的估计.
- 极大似然估计的构造思想非常简洁,是统计思想的最好体现. 在很多参数估计或者模型估计中，极大似然估计的思想被大量使用.

Cons:

- 极大似然估计的表现依赖于似然函数的选取,在实际数据中必须选择合适的总体分布(需要具体的形式).
- 极大似然估计中的优化问题有的时候没有显示解(影响可解释性)或者不容易计算.

回归中的MLE

假定应变量Y和解释变量X有线性关系：

$$Y_i = \mathbf{X}_i^\top \beta + \epsilon_i, \ \ i = 1, \ldots, n,$$

其中ϵ_i是噪音且$E\epsilon_i = 0, var(\epsilon_i) = \sigma^2$.

这时候没有具体的分布，没有办法写出似然函数，所以还需要额外的假设ϵ_i满足某个或者某类分布.

注意

使用极大似然估计必须有具体的分布来构造出似然函数！

正态噪音

假定$\epsilon_1, \cdots, \epsilon_n$, $i.i.d.\ \sim N(0, \sigma^2)$时候，我们有

$$Y_i \sim N(\mathbf{X}_i^\top \beta, \sigma^2),\ i = 1, \ldots, n.$$

我们可以写出似然函数

$$L(\beta, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(Y_i - \mathbf{X}_i^\top \beta)^2\},$$

正态噪音

假定$\epsilon_1, \cdots, \epsilon_n$, $i.i.d.\ \sim N(0, \sigma^2)$时候，我们有

$$Y_i \sim N(\mathbf{X}_i^\top \beta, \sigma^2),\ i = 1, \dots, n.$$

我们可以写出似然函数

$$L(\beta, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(Y_i - \mathbf{X}_i^\top \beta)^2\},$$

- 这里把σ^2也当未知参数放进了模型.
- 似然函数中逻辑上应该是$f(x, y, \beta, \sigma) = f(y|x)f(x)$，我们去除了对参数没有影响的$f(x)$部分.

极大似然估计为

$$\arg\max_{\beta\in\mathbb{R}^p,\sigma^2} \frac{1}{(\sqrt{2\pi}\sigma)^n}\exp\{-\frac{1}{2\sigma^2}(\mathbb{Y}-\mathbb{X}\beta)^\top(\mathbb{Y}-\mathbb{X}\beta)\},$$

如果只关注回归系数β部分，

$$\hat{\beta}=\arg\min_{\beta\in\mathbb{R}^p}(\mathbb{Y}-\mathbb{X}\beta)^\top(\mathbb{Y}-\mathbb{X}\beta).$$

当噪音为正态分布时，极大似然估计等价于最小二乘估计.

极大似然估计为

$$\arg\max_{\beta\in\mathbb{R}^p,\sigma^2} \frac{1}{(\sqrt{2\pi}\sigma)^n}\exp\{-\frac{1}{2\sigma^2}(\mathbb{Y}-\mathbb{X}\beta)^\top(\mathbb{Y}-\mathbb{X}\beta)\},$$

如果只关注回归系数β部分，

$$\hat{\beta} = \arg\min_{\beta\in\mathbb{R}^p}(\mathbb{Y}-\mathbb{X}\beta)^\top(\mathbb{Y}-\mathbb{X}\beta).$$

当噪音为正态分布时，极大似然估计等价于最小二乘估计.

思考

如果噪音不是正态分布，例如对称均匀分布、对称指数分布等，极大似然估计的形式是什么样的?

指数噪音

假定噪音是对称的指数分布，即 Laplace 分布：

$$\epsilon_i \sim \text{Laplace}(0, \frac{\sigma}{\sqrt{2}}) : f(x) = \frac{1}{\sqrt{2}\sigma} \exp\{-\frac{\sqrt{2}|x|}{\sigma}\},$$

这里$E\epsilon_i = 0, var(\epsilon_i) = \sigma^2$, 可得极大似然估计：

$$\arg\max_{\beta \in \mathbb{R}^p, \sigma^2} \frac{1}{(\sqrt{2}\sigma)^n} \exp\{-\frac{\sqrt{2}}{\sigma}\|\mathbb{Y} - \mathbb{X}^\top \beta\|_1\}$$

以及

$$\text{最小一乘法:} \quad \hat{\beta} = \arg\min \sum_{i=1}^{n} |Y_i - \mathbf{X}_i^\top \beta|.$$

Section 5

矩估计–最优线性投影

最优线性投影

把样本$(\mathbf{X}_1, Y_1), \cdots, (\mathbf{X}_n, Y_n)$看成是来自于总体$(\mathbf{X}, Y)$，考虑最优的线性投影：

$$\arg\min_{\beta \in \mathbb{R}^p} E(Y - \mathbf{X}^\top \beta)^2.$$

即从总体角度来找到最优的线性投影.

考虑一般的最优线性投影:

$$\arg\min_{\alpha\in\mathbb{R},\beta\in\mathbb{R}^p} E(Y-\alpha-\mathbf{X}^\top\beta)^2.$$

通过计算可以得到解为:

$$\begin{aligned}\beta^* &= \arg\min E((Y-EY)-(\mathbf{X}-E\mathbf{X})^\top\beta)^2\\ &= \arg\min\{\beta^\top \mathrm{cov}(\mathbf{X})\beta - 2\beta^\top \mathrm{cov}(\mathbf{X},Y)\}\\ &= \Sigma^{-1}\mathrm{cov}(\mathbf{X},Y), \qquad (\Sigma\text{严格正定条件下})\end{aligned}$$

以及

$$\begin{aligned}\alpha^* = \arg\min_{\alpha\in\mathbb{R}} E(Y-\alpha-\mathbf{X}^\top\beta^*)^2 &= E(Y-\mathbf{X}^\top\beta^*)\\ &= EY-(E\mathbf{X})^\top\beta^*.\end{aligned}$$

从最优线性投影的角度Σ或Σ^{-1}影响到投影角度.

协方差矩阵

在多元/高维数据分析中，总体协方差矩阵

$$\Sigma = cov(\mathbf{X}) = (\mathrm{cov}(X_i, X_j))_{p\times p},$$

是一个非常重要的参数，其刻画了数据的离散程度(类似于一元随机变量中的方差). 总体协方差矩阵的逆矩阵$\Omega = \Sigma^{-1}$也是一个重要的度量，一般称为精度矩阵(precision matrix)

备注

Σ和Ω是过去二十年高维统计的主要研究课题，是高斯图模型(Gaussian Graphical Model)中的核心参数.

在具体数据集中，一般会用样本协方差矩阵和样本协方差来估计Σ和$\mathrm{cov}(\mathbf{X}, Y)$:

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^\top,$$

$$\widehat{\mathrm{cov}(\mathbf{X}, Y)} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})(\mathbf{X}_i - \bar{\mathbf{X}}).$$

可以得到对应估计

$$\hat{\beta} = \hat{\Sigma}^{-1}\widehat{\mathrm{cov}(\mathbf{X}, Y)},\ \hat{\alpha} = \bar{Y} - \bar{\mathbf{X}}^\top\hat{\beta}.$$

备注

注意到最小二乘法中数据矩阵第一列是1，通过矩阵运算可以验证上述结果和$(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\mathbb{Y}$是完全一样的.

回归函数

在最优线性投影的基础上，我们考虑任意的函数$g(\cdot)$，

$$\arg\min E(Y - g(\mathbf{X}))^2,$$

记$Z = E(Y|\mathbf{X})$, 我们有

$$\begin{aligned}
&E(Y - g(\mathbf{X}))^2\\
=&E(Y - Z + Z - g(\mathbf{X}))^2\\
=&E(Y - Z)^2 + E(Z - g(\mathbf{X}))^2 + 2E(Y - Z)(Z - g(\mathbf{X}))\\
=&E(Y - Z)^2 + E(Z - g(\mathbf{X}))^2,
\end{aligned}$$

所以在均方损失下，最优的函数为$f(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$. 最小二乘估计是假设了回归函数是线性的.

Thank you!