《线性回归》
—（非参数）回归分析

杨　　瑛

清华大学　数学科学系
Email: yangying@mail.tsinghua.edu.cn
Tel: 62796887

2019.03.05

主要内容：（非参数）回归分析

1. 回归分析
   - 主要想法
   - 线性回归
   - 非参数回归
   - 离散的响应变量
   - 连续响应的非参数回归估计：最近邻估计
   - 局部平均
   - 局部平均的影响
   - 均方误差
   - R-lab

主要想法:

- 回归分析在于考察(单个)因变量Y和一个或多个自变量$\mathbf{X}_1, \cdots, \mathbf{X}_k$ 之间的关系。
- 回归分析描述给定$x_1, \cdots, x_k$ 后$\mathbf{Y}$ 的条件分布:

$$f(\mathbf{Y}|x_1, \cdots, x_k) = f(\mathbf{Y}|\mathbf{X}_1 = x_1, \cdots, \mathbf{X}_k = x_k).$$

  通常我们描述这个分布的均值或者中位数，即条件均值或者条件中位数。
- 它可以用于:
  - ✓ 描述$\mathbf{Y}$如何依赖于$\mathbf{X_1}, \cdots, \mathbf{X_k}$
  - ✓ 从$\mathbf{X_1}, \cdots, \mathbf{X_k}$预测$\mathbf{Y}$
  - ✓ 对$\mathbf{X_1}, \cdots, \mathbf{X_k}$关于$\mathbf{Y}$的效应做推断。

线性回归

- 全称: 普通最小二乘多重线性回归(Ordinary least squares multiple linear regression)。
- 线性回归的假设:
  - ✓ 数据对感兴趣的总体具有代表性。
  - ✓ $\mathbf{E}[Y|\mathbf{x_1},\cdots,\mathbf{x_k}]$ 是$\mathbf{x_1},\cdots,\mathbf{x_k}$ 的线性函数，$\mathbf{E}[Y|\mathbf{x_1},\cdots,\mathbf{x_k}]=\sum_{i=1}^{k}\theta_i\mathbf{x}_i$
  - ✓ $f(\mathbf{Y}|\mathbf{x_1},\cdots,\mathbf{x_k})$ 的方差不依赖于$\mathbf{x_1},\cdots,\mathbf{x_k}$。
  - ✓ $f(\mathbf{Y}|\mathbf{x_1},\cdots,\mathbf{x_k})$ (近似)正态。

线性回归

- 假定有观测值$(y_i, \mathbf{x}_{i1}, \cdots, \mathbf{x}_{ik})$,$i = 1, \cdots, n$, 建立模型

$$y_i = \sum_{j=1}^{k} \theta_j \mathbf{x}_{ij} + \epsilon_i, i = 1, \cdots, n.$$

- 可以用**最小二乘法**来估计未知参数$\theta = (\theta_1, \cdots, \theta_k)$.
- 其它方法。（MLE?）
- 可以分别讨论$k = 2$和$k > 2$的情形。更多细节将在后面的课程中逐步展开，

非参数回归

- 非参数回归$\mathbf{E}[\mathbf{Y}|\mathbf{X}=\mathbf{x}]=m(\mathbf{x})$, $m(\cdot)$是未知的函数。
- 为了简单起见，考虑$\mathbf{Y}$和$\mathbf{X}$都是一维的情形。
  假定有观测值$(y_i,\mathbf{x}_i)$, $i=1,\cdots,n$, 建立模型

$$y_i=m(\mathbf{x}_i)+\epsilon_i, i=1,\cdots,n, \tag{1}$$

  其中$m(\cdot)$是定义在区间$I=[a,b]$上的函数，$\mathbf{E}[\epsilon_i]=0$ 或者median$(\epsilon_i)=0$.
  模型(1)是所谓的**非参数**（均值或者中位数）**回归模型**。

- ♠ 非参数回归不假设线性和正态性等, 非常灵活。
  【画示意图】
- ♠ 为什么考虑呢?
  - ✓ 非常弱的假设
  - ✓ 也有局限性
  - ✓ 非参数回归的现代方法正在兴起

离散的响应变量

- 回顾: 回归分析描述了条件分布$f(\mathbf{Y}|\mathbf{x_1},\cdots,\mathbf{x_k})$。
- 在非常大的样本中，如果**X**是离散的，我们可以直接检查这个条件分布。
- 但是如果有很多的协变量，要搞清楚条件分布就成了问题：
  - ✓ 三个具有10种可能结果的协变量可以给出$10^3=1000$种组合。
  - ✓ 要得到分布的信息，需要在每一种情形之下有足够的数据，这就需要一个非常庞大的数据集。
  - ✓ 这就是所谓的"维度祸根"(curse of dimensionality)。
  - ✓ 我们通常考虑**X**的维数是一维或者二维的非参数回归。

连续响应的非参数回归估计：最近邻估计

- 回顾：假定有观测值$(y_i, \mathbf{x}_i)$, $i = 1, \cdots, n$, 建立模型

$$y_i = m(\mathbf{x}_i) + \epsilon_i, i = 1, \cdots, n, \tag{2}$$

其中$m(\cdot)$是定义在区间$I = [a, b]$上的函数，$\mathbf{E}[\epsilon_i] = 0$.

- 如何利用观测值估计未知的函数$m(x), x \in [a, b]$. 即，对于没一个$x \in [a, b]$, 要得到$m(x)$ 的估计$\widehat{m}(x)$.

连续响应的非参数回归估计：最近邻估计

- 解决方案之一:
  - ✓ 任意给定$x \in (a, b)$和$h > 0$, 做x的邻域$I(x, h) = (x - h, x + h)$, $m(x)$的估计可以定义为：

$$\widehat{m}_h(x) = \frac{\sum_{i:\mathbf{x}_i \in I(x,h)} y_i}{\#\{\mathbf{x}_i : \mathbf{x}_i \in I(x, h)\}}, \tag{3}$$

    即，为估计函数$m(x)$在x处的值，将x的一个邻域内的观测值做平均即可。
  - ✓ 实际做的时候，需要将区间$[a, b]$分为若干个小区间，在每个小区间中选择一个代表点，比如中点作为上面的x.
  - ✓ 上面定义的估计合理吗？如果h太小，有可能区间$I(x, h)$中没有观测值！

连续响应的非参数回归估计：最近邻估计

- 解决方案之二:
  - ✓ 任意给定$x \in (a,b)$和$h \in N$, $I(x,h)$中包含了离x最近的h个观测值, 即将$|x_i - x|, i = 1, \cdots, n$从小到大排序，对于任意规定的$x$, $x_{(1)}(x), \cdots, x_{(n)}(x)$为$x_1, \cdots, x_n$的一个排序，$y_{(i)}(x)$为$x_{(i)}(x)$处对应的观测值。$m(x)$的估计可以定义为：

$$\widehat{m}_h(x) = \frac{1}{h}\sum_{i=1}^{h} y_{(i)}(x), \tag{4}$$

    即，为估计函数$m(x)$在x处的值，将与x最近的h个观测值做平均即可。
  - ✓ 上面定义的估计合理吗？

局部平均

- 上面给出的方法(3)和(4)都是所谓的局部平均方法。
- (3)和(4)中的x在实际中可以选择为$x_1, \cdots, x_n$, 即估计$m(x)$在观测点处的值即可。
- 局部平均带条的方法非常粗糙。我们只在很少的点上做了估计
- 解决方案: 使用重叠带条(移动窗口):
  - ✓ 使用每一个x的值作为中点
  - ✓ 使用固定宽度窗口或包含固定数量数据点的窗口

局部平均的影响

- 前几个局部平均值和后几个局部平均值是相同的
- 线是粗略的-如果观察进入和离开窗口，平均值会上下跳动
- 异常数据值(异常值)会产生很大的影响

  我们可以通过加权来解决第二个和第三个问题:
- 对靠近窗口中心的观测给予较大的权重，对靠近窗口边缘的观测给予较小的权重
- 对边远的观测值给予较小的重视

  这一方法以及其它的一些方法，被内置到R的loess平滑器中

  散点图通常有助于看到数据中的模式
- 最近邻估计或者loess得到的结果有助于初步直观的了解两个变量之间的关系，为进一步提出$\mathbf{E}[\mathbf{Y}|\mathbf{X}=x]=m(x)$的参数模型的假设提供依据和诊断。

均方误差

- 为评价估计$\widehat{m}_h(x)$的行为，我们考虑其均方误差

$$\text{MSE}(\widehat{m}_h(x)) = \mathbf{E}[\widehat{m}_h(x) - m(x)]^2. \tag{5}$$

很容易得到：

$$\begin{aligned}\text{MSE}(\widehat{m}_h(x)) &= \text{Var}[\widehat{m}_h(x)] + (\mathbf{E}[\widehat{m}_h(x)] - m(x))^2 \\ &= V(x,h) + (\text{Bias}(x,h))^2\end{aligned} \tag{6}$$

这就是《统计推断》中的bias-variance分解。

均方误差

- 对于任意固定的x, 我们可以

$$\min_{h\in H_n} \mathsf{MSE}(\widehat{m}_h(x)), \tag{7}$$

其中H_n为指标集。
在方差和偏差之间达到折中的平衡以确定最优的h. 可以看出，由(7)确定的最优的$h^\star$为

$$h^\star = h^\star(x) = h^\star(x, n),$$

注意这些符号的意思。

均方误差

- 为了更加明确的研究估计(4)的性质(6)和(7), 我们考虑更加具体的模型：

$$y_i = m(x_i) + \epsilon_i, i = 1, \cdots, n, \tag{8}$$

其中$x_i = \frac{i}{n}$, $\epsilon_1, \cdots, \epsilon_n$ iid, $\mathbf{E}[\epsilon_i] = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2 < \infty$. 其中$H_n$为指标集。

- 在模型(8)下，可以推导估计(4)的性质(6).

R-lab

- 具体内容和要求在课堂上布置。（估计结果可以与loess的结果进行比较）