第三讲：一元线性回归及均值回归

软件工程《机器学习》课程

李济洪，王瑞波

山西大学 • 软件学院

wangruibo@sxu.edu.cn

2019 年 9 月 17 日

提纲

统计学习算法

算法一般形式

$$y = f(x) + \epsilon \quad (1)$$

$$\hat{y} = \hat{f}(x) \quad (2)$$

基本概念

- 预测变量：y；
- 响应变量：$\hat{y}$；
- 统计学习方法：f；
- 训练数据集:
  $D_n = \{z_i\}_{i=1}^n = \{(x_i, y_i)\}$；

统计学习算法建模过程

1. 选取算法族，抽象出算法形式；
   1. 本课程主要考虑参数算法；
2. 基于训练数据，估计算法参数；
3. 根据算法参数估计，基于算法的参数形式，对未来的观测进行预测。

算法比较涉及的问题

1. 如何评估一个统计学习算法的性能？
2. 如何对比多个统计学习算法的性能？
3. 如何解释预测变量与响应变量之间的关系？

本讲内容

- 一元线性回归算法
- 均值回归算法

回顾：回归

- 回归：响应变量为连续值；

示例

Table: Advertising 数据集示例

Id	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
		.....		
200	232.1	8.6	8.7	13.4

思考

1. 电视媒体的预算与产品销量之间有关系么？
2. 它们是什么关系？关系有多强？
3. 如何精确地估计电视媒体的预算与产品销量之间的关系？
4. 给定未来电视媒体的预算，如何预测未来销量？该预测的精度如何？

电视媒体的预算与产品销量之间的关系

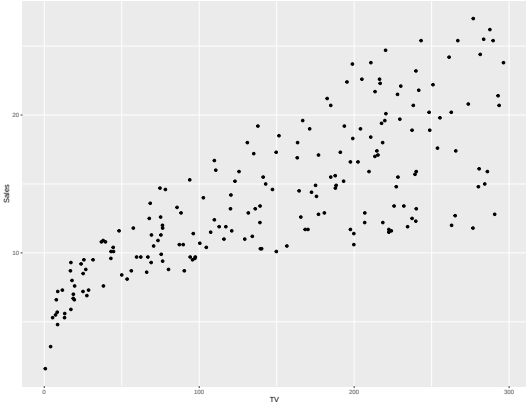


Figure: TV 预算与销量间关系

观察得到的结果：

- TV 预算越大、销量越多；
- “似乎”成线性增长；

基于上述几点观察，可首先考虑采用线性统计方法进行建模。

$$\text{Sales} = \beta_0 + \beta_1 \times \text{TV} \tag{3}$$

回看四个问题

$$\text{Sales} = \beta_0 + \beta_1 \times \text{TV}$$

四个问题：

1. 电视媒体的预算与产品销量之间有关系么？
2. 它们是什么关系？关系有多强？
3. 如何精确地估计电视媒体的预算与产品销量之间的关系？
4. 给定未来电视媒体的预算，如何预测未来销量？该预测的精度如何？

解答

1. 若 $\beta_1 \neq 0$, 则认为它们有关系。
2. 目前认为是线性关系，强度取决于 β_1 的大小。
3. 问题可转化为如何估算 β_0 和 β_1。
4. 基于 β_0 和 β_1 的估计后，可直接带入线性方程，对未来的 TV 值，计算出对应的 Sales 值。精度需要在未来的数据点上进行测量。

一元线性回归算法

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X} + \epsilon \tag{4}$$

$$\hat{\mathrm{Y}} = \hat{\beta}_0 + \hat{\beta}_1 \mathrm{X}_{\text{future}} \tag{5}$$

其中，$\epsilon \sim \mathrm{N}(0, \sigma^2)$。

基本概念

- 预测变量：X；
- 响应变量：Y（连续值）；
- 响应变量的预测值：$\hat{\mathrm{Y}}$
- 模型误差：ϵ；
- 模型系数或参数：β_0 和 β_1；
- 参数估计：$\hat{\beta}_0$ 和 $\hat{\beta}_1$；

一元线性回归算法的等价形式

$$\mathrm{Y}|\mathrm{X} \sim \mathrm{N}(\beta_0 + \beta_1 \mathrm{X}, \sigma^2). \tag{6}$$

显然：

$$\mathrm{E}[\mathrm{Y}] = \beta_0 + \beta_1 \mathrm{X} \tag{7}$$

也就是说，我们希望找到的趋势是关于 Y 的均值的趋势。具体地，我们希望使用一条直线来拟合响应变量 Y 的均值。

一元线性回归算法图解

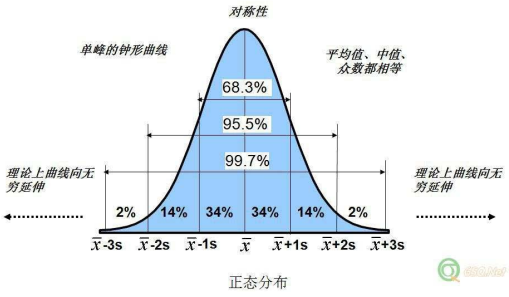


Figure: 正态分布密度函数

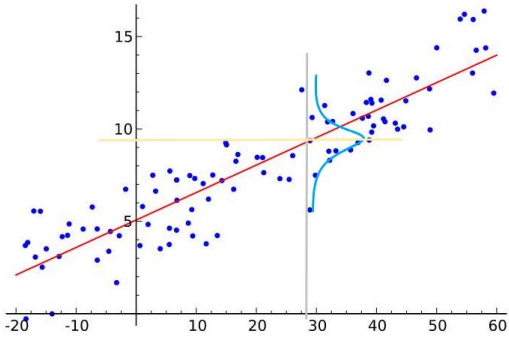


Figure: 回归函数图示

关于正态分布

- 最最最常用的连续分布。
- 该分布有很好的理论性质和物理解释。
- 所有的概率统计交叉都有相关介绍。
- 正态分布的发现历史，可参考：
  - 《正态分布的前世今生》(https://songshuhui.net/archives/76501)

无截距项的一元线性回归

截距项

参数 β_0 被称为截距项。

无截距项的一元线性回归算法

$$Y = \beta_1 X + \epsilon$$

- 加入人为约束（假设）：当 $X = 0$ 时，Y 必须等于 0；
  - 该约束很可能不成立。
- 对应的回归线图像需经过原点。
- 为一般一元线性回归算法在 $\beta_0 = 0$ 时的特殊情形。（不做单独讨论！）

参数估计的问题设立

优化问题

给定训练数据 $\mathrm{D_n} = \{(\mathrm{x_i}, \mathrm{y_i})\}_{\mathrm{i=1}}^{\mathrm{n}}$，如何估计一元线性回归算法中的参数 β_0 和 β_1，使所得到的模型可以更“贴近”训练数据？

何为更贴近？

所谓“贴近”，可认为训练数据应遵循残差平方和最小准则：

$$\mathrm{RSS} = \sum_{\mathrm{i=1}}^{\mathrm{n}} [\mathrm{y_i} - (\beta_0 + \beta_1 \mathrm{x_i})]^2 \tag{8}$$

参数估计的优化问题

$$\hat{\beta}_0, \hat{\beta}_1 = \arg\min_{\beta_0, \beta 1} \sum_{\mathrm{i}=1}^{\mathrm{n}} [\mathrm{y_i} - (\beta_0 + \beta_1 \mathrm{x_i})]^2 \tag{9}$$

思考：如何求解？

解 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的形式

$$\hat{\beta}_1 = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{x_i} - \bar{\mathrm{x}})(\mathrm{y_i} - \bar{\mathrm{y}})}{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{x_i} - \bar{\mathrm{x}})^2} \tag{10}$$

$$\hat{\beta}_0 = \bar{\mathrm{y}} - \hat{\beta}_1 \bar{\mathrm{x}} \tag{11}$$

其中，$\bar{\mathrm{x}}$ 和 $\bar{\mathrm{y}}$ 为训练数据上的均值。该估计值被称为一元线性回归的**最小二乘估计**。

回看广告数据例子：$\text{Sales} = 7.03 + 0.0475 \times \text{TV}$

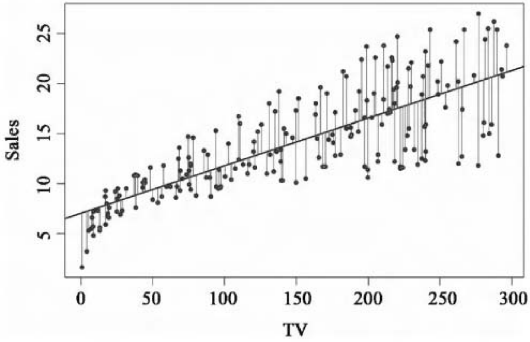
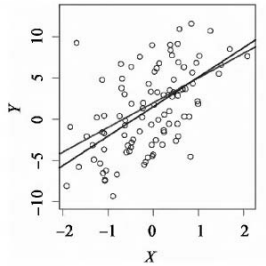


Figure: 广告数据集上的回归函数

回归直线

- 总体回归直线：β_0 和 β_1 的真值对应的回归直线。
  - 可看作数据总体上的回归线（理论值、理想值）。
  - 例如：$\mathrm{Y} = 2 + 3\mathrm{X} + \epsilon$
- 最小二乘线：基于训练数据 $\mathrm{D_n}$，得到的回归线。
  - n 是有限的。
  - $\mathrm{D_n}$ 不同，回归线也不同。例如第 1 组：$\beta_0 = 2.1, \mathrm{beta}_1 = 3.4$;



最小二乘线的准确性

何为“准确性”

“最小二乘线”是否很好地靠近了“总体回归直线”？

如何刻画最小二乘线的“准确性”

1. 两方面刻画：
   - 最小二乘线偏离总体回归直线多大？
   - 随着 D_n 变化，最小二乘线的变化是否剧烈？
2. 问题转化：$\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的准确性。
   - $\hat{\beta}_i$ 与真值 β_i 之间的偏差多大？
   - 随 D_n 变化，$\hat{\beta}_i$ 的方差多大？

参数估计 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的准确性

参数估计的 MSE

$$\mathrm{MSE}(\hat{\beta}_\mathrm{i}) = \mathrm{Bias}^2(\hat{\beta}_\mathrm{i}) + \mathrm{Var}(\hat{\beta}_\mathrm{i}) \tag{12}$$

其中，$\mathrm{i} = 0, 1$。

- $\mathrm{Bias}(\hat{\beta}_\mathrm{i}) = \mathrm{E}[\hat{\beta}_\mathrm{i}] - \beta_\mathrm{i}$ 为参数估计 $\hat{\beta}_\mathrm{i}$ 的偏差。
- $\mathrm{Var}(\hat{\beta}_\mathrm{i})$ 为参数估计 $\hat{\beta}_\mathrm{i}$ 的方差。

参数估计 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的偏差

最小二乘估计的偏差

$$\mathrm{E}[\hat{\beta}_{\mathrm{i}}] - \beta_{\mathrm{i}} = 0 \tag{13}$$

即：最小二乘估计为无偏估计。

参数估计 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的方差

$$\mathrm{Var}[\hat{\beta}_0] = \sigma^2 \left[\frac{1}{\mathrm{n}} + \frac{\bar{\mathrm{x}}^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{x_i} - \bar{\mathrm{x}})^2} \right] \tag{14}$$

$$\mathrm{Var}[\hat{\beta}_1] = \frac{\sigma^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{x_i} - \bar{\mathrm{x}})^2} \tag{15}$$

- 方差值与 n 有关，当 $\mathrm{n} \to \infty$，方差均趋于 0.
- **理论结论**：最小二乘估计是方差最小的无偏估计。
- **标准误**：$\mathrm{SE}(\beta_\mathrm{i}) = \sqrt{\mathrm{Var}[\beta_\mathrm{i}]}$，其中，$\mathrm{i} = 1, 2$。

参数估计 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的置信区间

置信区间

置信区间展现的是这个参数的真实值有一定概率落在测量结果的周围的程度，其给出的是被测量参数的测量值的可信程度，即前面所要求的“一个概率”。

$$\mathrm{CI}_{95\%}(\hat{\beta}_\mathrm{i}) = [\hat{\beta}_0 - 2\mathrm{SE}(\hat{\beta}_\mathrm{i}), \hat{\beta}_0 + 2\mathrm{SE}(\hat{\beta}_\mathrm{i})] \tag{16}$$

其中，$\mathrm{i} = 1, 2$。

思考

为何要将标准误乘以 2？

预测变量与响应变量是否有关？

假设检验

H_0 : X 与 Y 之间没有关系 v.s. H_1 : X 与 Y 之间没有关系；

在一元线性回归中，该问题退化为如下检验问题：

参数的假设检验

$H_0 : \beta_1 = 0$ v.s. $H_1 : \beta_1 \neq 0$；

难点：

无法直接观测到 β_1 的值，只能计算其估计 $\hat{\beta}_1$。如何根据后者推断 β_1 的性质？

检验统计量

检验问题的解决思路

依赖于 $\hat{\beta}_1$ 的准确性，即 $\mathrm{SE}(\hat{\beta}_1)$。

t-检验统计量：

$$\mathrm{T} = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)} \sim \mathrm{t}_{0.95}(\mathrm{n}-2) \tag{17}$$

其中，$\mathrm{t}_{0.95}(\mathrm{n}-2)$ 可查分位数表。

若 $\mathrm{T} > \mathrm{t}_{0.95}(\mathrm{n}-2)$, 则拒绝 H_0，认为 X 与 Y 之间存在关系；否则，无法拒绝掉 H_0，意味着当前的数据无法表明 X 与 Y 之间存在关系。

p

回看广告数据集例子

Table: 广告数据参数的准确性

	截距项 β_0	参数 β_1
系数	7.0325	0.0475
标准误	0.4578	0.0027
t 统计量	15.36	17.67
p 值	< 0.0001	< 0.0001

模型的准确性评价

回顾建模的两个目标

1. 推断：模型对训练数据的拟合程度有多好？
2. 预测：模型在未来数据上的表现有多好？

一元线性回归的准确性

可通过如下两个量进行评价：

1. 残差标准误；
2. 拟合优度 R^2 统计量；

残差标准误

RSE(Residual Standard Error).

$$\hat{\sigma}^2 = \mathrm{RSE} = \sqrt{\frac{\mathrm{RSS}}{\mathrm{n}-2}} = \sqrt{\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}[\mathrm{y_i} - (\hat{\beta}_0 + \hat{\beta}_1 \mathrm{x_i})]^2}{\mathrm{n}-2}} \tag{18}$$

- RSE 值越小，代表模型越好地拟合了训练数据。
- RSE 取值范围为 $[0, +\infty)$。

拟合优度 R^2 统计量

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}. \tag{19}$$

释义

- TSS: 总平方和

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{20}$$

- RSS：回归平方和
- **理论结果**：$0 < RSS < TSS$，$R^2 \in [0, 1]$。

R^2 的解释

R^2 的特点

1. $R^2 \in [0,1]$ 比 RSE 有更好的解释性。
2. R^2 越大，说明模型拟合越好。

与相关系数的关系

在一元线性回归模型中，下式成立：

$$R^2 = Cor(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{21}$$

广告数据集上的模型准确性

- 残差标准差：3.26
- 拟合优度 R^2：0.612

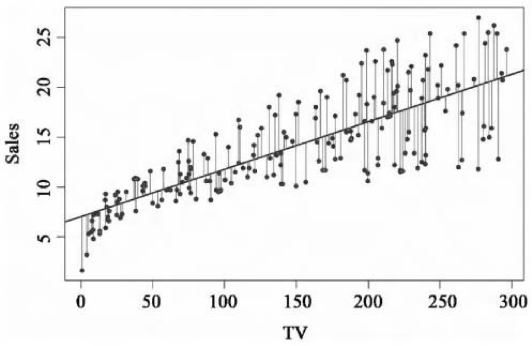
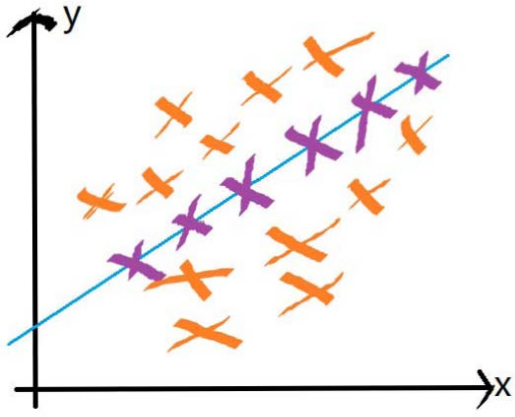


Figure: 广告数据集上的回归函数

如何从预测角度评判模型的准确性？

思考

- RSE 和 R^2 只揭示了算法在训练集上的性能？
  - 可能存在过拟合问题。
- 如何从预测角度评判模型的准确性？



Github 讲解

- git 的基本用法;
  - 分布式的含义
  - clone pull push checkout commit(checkin)
- github 的一些知识;
- Microsoft 与 github 之间的关系;
  - 软件产品
  - 软件创作
- Github Desktop 的使用;
  - 新建 statLearn 项目;

数据集的读取

- 观察 advertising 数据集的格式；
- ‘read.csv‘函数和 ‘read.table‘函数的作用；
- 文件的绝对路径对程序移植性的影响；
- 全局变量及 ‘Sys.getenv()‘函数的用法；

Rstudio 开发环境讲解

- R-CRAN 包的基本结构。以 R-CRAN 中的任意包为例（https://cran.r-project.org/）
- Rstudio 的项目概念。
- 工作目录的读取及设置。

回顾统计学习算法的建模流程

1. 读取实验数据；
2. 观察时间数据，选取合适的统计学习模型；
3. 训练统计学习模型；
4. 评估统计学习模型的解释性（推断）；
5. 使用统计学习模型进行预测；
6. 评估统计学习的预测性能（预测）；

R 工程的目录规划

1. 存放数据集的目录: datasets
2. 存放机器学习算法的目录: algorithms
3. 任务实例目录: impls

规划数据集格式配置文件

1. 数据集的文件名: 贴切、有意义；
2. 数据集中代码结构规划：
   - ‘DataGenerate(dataConf)‘函数
   - ‘Prepackages‘向量
   - ‘Validate(dataConf)‘函数;
3. 数据集配置: ‘dataConf‘。该配置为 list 结构。

数据集格式

- 最后一列为响应变量；第一列至倒数第二类为预测变量。
- 存储格式为 data.frame。

观察实验数据

常用画图函数

plot(advertising.dataTV, advertising.dataSales)

ggplot

ggplot(advertising.data, aes(TV, Sales))+ geom_point()

规划统计学习算法文件

1. 算法文件的文件名：贴切、有意义
2. 算法文件中代码结构规划:
   - ‘train‘函数
   - ‘predict‘函数
   - ‘assess‘函数
   - validation 函数
3. 算法配置：‘algorConf‘。该配置为 list 结构。

数据文件和算法配置文件的 validate 函数实现

- 数据集的配置：指定所用的预测变量；
- 算法的配置：有无截距项；

算法文件中各个函数的实现

train 函数实现

方式一：根据前面所学的理论知识，给出 $\hat{\beta}_0$ 和 $\hat{\beta}_1$。

方式二：使用 lm 函数，对线性回归算法进行建模。

predict 函数实现

方式一：自己编写预测代码；

方式二：调用 predict 函数

assess 函数实现

方式一：自己编写预测代码；

方式二：调用 summary 函数

本讲作业

1. 基于残差平方和准则，给出一元线性回归算法的参数 β_0 和 β_1 的估计 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的详细求解过程。
2. 思考：从预测角度，应该如何评价一元线性回归算法的准确性？请给出具体思路。

谢谢！

Questions & Answering！