ZOE's MindMap

https://woolelf.github.io/

14-17 | v1.0 | 20170320

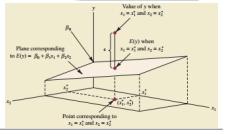
多元回归

模型基础

- 多元回归模型
  - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$
  - 描述 y 是如何依赖 → 自变量和误差项 → 的数学方程
  - 误差项 → 包含在 y 里面但不能被 p 个自变量的线性关系 → 解释的变异性
- 多元回归方程
  - $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$
  - 应变量 y 的「期望值」与自变量的值 → 之间关系的数学方程
- 估计的多元回归方程
  - $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$ → 解释 → $b_0, b_1, b_2, \ldots, b_p$ 给出了 → $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ 的估计值
  - 求解 → 最小二乘法则 → $\min \Sigma(y_i - \hat{y}_i)^2$
  - 根据样本数据和最小二乘法 → 得到的 → 多元回归方程的估计
- 回归系数 → 解释
  - 当所有其他自变量 → 都保持不变时
  - b_j 含义 → 对应自变量 x_j → 变化一个单位时
    - 应变量 y 变化程度的一个估计
- 多元判定系数
  - 定义
    - $R^2 = \frac{SSR}{SST}$ → $SST = SSR + SSE$
    - 对估计的多元回归方程「拟合优度」的度量
    - * 100 → 应变量 y 中的「变异性」能被 → 估计的多元回归方程解释的百分比
  - ★ 修正多元判定系数
    - 方法 → 用自变量个数 → 进行调整
    - 理由
      - 增加「自变量」 → 将影响估计回归方程解释的百分比
      - R^2 会变大
      - 目的 → 避免过高估计 → 由于自变量数目增加 → 而带来的影响
    - 公式
      - $R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$
      - n → 观测值的数目
      - p → 自变量的数目
    - 注意
      - R^2 值较小
      - 自变量数目较多
      - → 修正多元判定系数为负 → 最后取值为0

模型的假定

- ϵ 的假定
  - 同简单线性回归
    - 1. 平均值 / 期望值为 0 → $E(\epsilon) = 0$
    - 2. 对于所有 x 值 → ϵ 的方差均相同 → 记为 σ^2 → y 关于「回归直线」的方差等于 σ^2
    - 3. → ϵ 的值是「相互独立」的
    - 4. → ϵ 是一个「正态分布」的随机变量
      - y 也是一个「正态分布」的随机变量
      - → ϵ 是 y 值和 y 的期望值之间的离差
  - $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ → 图示



显著性检验

- 对比
  - 简单线性回归 → t 检验和 F 检验 → 目的与结论相同
  - 多元回归
    - F 检验
      - 总体的显著性检验
      - 目的 → 应变量和所有自变量之间 → 是否存在显著性关系
    - t 检验
      - 单独显著性检验
      - 目的
        - F 检验 → 已经表明「模型总体」的显著性
        - t 检验 → 用于确定每一个「单个的自变量」是否显著
        - 即 → 应变量和单个自变量之间 → 是否存在显著关系
- ▶ F 检验
  - 总体显著性的 F 检验
  - 建立假设
    - H_0: → $\beta_1 = \beta_2 = \cdots = \beta_p = 0$
    - H_a: → 至少有一个参数不等于 0
  - 检验统计量
    - $F = \frac{MSR}{MSE}$
      - $MSR = \frac{SSR}{p}$ → H0 → 成立 → σ^2 的无偏估计量
      - 不成立 → 高估 σ^2
      - $MSE = \frac{SSE}{n-p-1}$ → 无论 H0 是否成立 → σ^2 的无偏估计量
      - H0 成立时，F 值接近 1，否则 F 值将变大
  - 拒绝规则
    - Reject H_0 if p-value $\le \alpha$
    - Reject H_0 if $F \ge F_\alpha$

Source	Sum of Squares	Degrees of Freedom	Mean Square	F
Regression	SSR	p	$MSR = \frac{SSR}{p}$	$F = \frac{MSR}{MSE}$
Error	SSE	$n-p-1$	$MSE = \frac{SSE}{n-p-1}$	
Total	SST	$n-1$		

- ▶ t 检验
  - 单个参数显著性的 t 检验
  - 建立假设 → 对于任意参数 β_i → $H_0: \beta_i = 0$ / $H_a: \beta_i \ne 0$
  - 检验统计量 → $t = \frac{b_i}{s_{b_i}}$
  - 拒绝规则 → Reject H_0 if $t \le -t_{\alpha/2}$ or if $t \ge t_{\alpha/2}$ → 自由度 $n-p-1$

Logistic 回归

- 应变量 y 仅有两个假定值的情形
  - 0
  - 1
- Logistic 回归方程
  - $E(y) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}$
  - 概率形式
    - $E(y) = P(y=1|x_1, x_2, \ldots, x_p)$
    - 自变量一组值已知时
    - $E(y)$ 给出了 y = 1 的概率
  - $E(y)$
    - 取值范围 → 0 - 1
    - S 形曲线
- 估计的 logistic 回归方程
  - $\hat{y} = \text{estimate of } P(y=1|x_1, x_2, \ldots, x_p) = \frac{e^{b_0+b_1 x_1+b_2 x_2+\cdots+b_p x_p}}{1+e^{b_0+b_1 x_1+b_2 x_2+\cdots+b_p x_p}}$
  - $\hat{y}$ 给出了 y = 1 的概率 → 自变量一组特定值已知时
- 显著性检验
  - ▶ 总体
    - $H_0: \beta_1 = \beta_2 = 0$
    - H_a: → 至少有一个参数不等于 0
    - G 统计量
      - 服从 → 抽样分布 → χ^2 分布
      - 自由度 = 模型中自变量个数
  - ▶ 单个自变量
    - $H_0: \beta_i = 0$
    - $H_a: \beta_i \ne 0$
    - $z_i = b_i / s_{b_i}$
    - 如何使用?
      - 表明 → 模型总体是著 → G 检验 → 首先
      - 确定 → z 检验 → 再用
      - 是否有显著作用 → 单个自变量对模型总体
- ★ 解释 Logistic 回归
  - 有利于一个事件发生的「机会比」
    - 定义 → 「事件将要发生的概率」与「该事件将不会发生的概率」的比
    - 公式 → $\text{odds} = \frac{P(y=1|x_1, x_2, \ldots, x_p)}{P(y=0|x_1, x_2, \ldots, x_p)} = \frac{P(y=1|x_1, x_2, \ldots, x_p)}{1-P(y=1|x_1, x_2, \ldots, x_p)}$
    - 对「机会比」的影响 → 当一组自变量中「只有一个自变量」增加了一个单位时
  - 机会比率
    - 定义
    - 公式 → $\text{Odds Ratio} = \frac{\text{odds}_1}{\text{odds}_0}$
  - 对数机会比（logit）变换
    - 来源 → 有利于 y=1 的机会比的自然对数 → $\ln(\text{odds}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$
    - 是自变量的线性函数
    - $g(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$
    - 对数机会比
      - 改写 → $E(y) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}$
      - logistic 回归方程
    - 估计的对数机会比
      - $\hat{g}(x_1, x_2, \ldots, x_p) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$

残差分析

- ▶ 标准化残差
  - $s_{y_i - \hat{y}_i} = s\sqrt{1-h_i}$
  - $\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}}$
  - 标准化残差图 → 正态分布假定 → 检测
  - 正态概率图
- ▶ 检测异常值
  - 学生化删除残差
    - 异常值 → 用来检测 → 一种标准化残差
    - 原理
      - 第 i 次观测值 → 1. 从数据集删除
      - 其余的 n-1 次观测值 → 2. 判用
      - 建立新的估计回归方程
      - 新（修正）的估计的标准误差 $s_{(i)}$ → 3. 计算
      - 第 i 次观测值为异常值 → $s_{(i)}$ 小于 s → 4. 如果
  - 两种识别方法 → ▶ 有影响的观测值
    - 杠杆率
      - 定义 → 反映自变量数值与平均值之间差距的度量
      - 经验法则 → $h_i > 3(p+1)/n$ → 杠杆率临界值
    - 库克距离测度
      - 定义
        - 第 i 次观测的杠杆率
        - 的基础上 → 建立在
        - 第 i 次观测残差
        - 观测影响程度 → 度量
        - 对估计回归方程有较强的影响 → 不一定 → 高杠杆率的观测值
      - 公式 → $D_i = \frac{(y_i - \hat{y}_i)^2}{(p+1)s^2}\left[\frac{h_i}{(1-h_i)^2}\right]$
        - 自变量的个数 → p
        - 估计的标准误差 → s
      - 第 i 次观测值是一个有影响的观测值 → $D_i > 1$ → 经验法则

分类自变量

- 含有「分类变量」的自变量
  - 定义
    - 分类自变量的影响的变量 → 用来模拟
    - 分类变量有 k 个水平
  - 虚拟变量 / 指标变量
    - 使用
      - 需要定义 k-1 个虚拟变量
      - 只能取值 1 或者 0
    - e.g.
      - 有 A，B，C 三个水平 → 销售区域 → 1. 分类变量
      - 2. 定义虚拟变量
        - $x_1 = \begin{cases} 1 \text{ if sales region B} \\ 0 \text{ otherwise} \end{cases}$
        - $x_2 = \begin{cases} 1 \text{ if sales region C} \\ 0 \text{ otherwise} \end{cases}$
      - 3. 建立回归方程
        - $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
        - $E(y \mid \text{region A}) = \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0$
        - $E(y \mid \text{region B}) = \beta_0 + \beta_1(1) + \beta_2(0) = \beta_0 + \beta_1$
        - $E(y \mid \text{region C}) = \beta_0 + \beta_1(0) + \beta_2(1) = \beta_0 + \beta_2$
      - 4. 三种变化情况
        - $x_1 = 0, x_2 = 0$ → A
        - $x_1 = 1, x_2 = 0$ → B
        - $x_1 = 0, x_2 = 1$ → C
        - 即

★ 多重共线性

- 定义 → 自变量之间的相关性
- 自变量之间不一定是独立的
  - 可能对模型估计产生问题
    - 任一自变量对应变量的单独影响 → 无法确定
    - 自变量高度相关时
  - 即
    - F 检验 → 显著
    - 单独参数 t 检验 → 却不显著
    - 样本相关系数绝对值大于 0.7 → ★ 对于两个自变量情形 → 多重共线性潜在影响 → 可能产生
    - 包含高度相关的自变量 → 模型中 → 尽可能回避 → 应用中