简单线性回归

https://woaielf.github.io/
—— made by ZY

1 回归模型

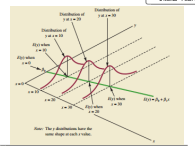
- 概念
  - 自变量 — 用于「预测或解释」的一个或多个变量 — x
  - 应变量 — 被预测或被解释的变量 — y
  - 回归模型 — 描述 y 如何依赖于 x、误差项 ϵ — 的方程
  - 回归方程 — 描述 y 的期望值 E(y) 如何依赖于 x 的方程
  - 估计的回归方程 — 根据样本数据 — 利用「最小二乘法」 — 建立「回归方程的估计」
  - ★ 注意 — 回归分析
    - 不能说明 — 因果关系
    - 只能表明 — 变量是如何（以怎样的程度）联系在一起的
- 简单线性回归
  - 定义
    - 含有一个自变量和一个应变量
    - 两个变量之间的关系用「一条直线」近似 — 回归分析
  - ▶ 简单线性回归模型 — $y = \beta_0 + \beta_1 x + \epsilon$
    - β_0 and β_1 — 模型参数
      - 截距
      - 斜率
    - ϵ — 模型误差项
      - 本质上 — 随机变量
      - 说明了 — 包含在 y 里面但「不能」被 x 和 y 之间的「线性关系」解释的「变异性」
  - ▶ 简单线性回归方程 — $E(y) = \beta_0 + \beta_1 x$ — 另一种形式 — $E(y|x) = \beta_0 + \beta_1 x$ — 强调 — 对于「x 的一个给定值」回归方程给出「y 的平均值」
  - ▶ 估计的简单线性回归方程 — $\hat{y} = b_0 + b_1 x$
    - b_0 and b_1 — 给出了 — β_0 and β_1 的估计值
    - $\hat{y}$
      - y 的平均值 E(y) 的一个点估计
      - y 的一个个别值的预测值
- ★ 最小二乘法
  - 目的 — 利用「样本数据」建立「估计的回归方程」
  - 准则 — $\min \Sigma(y_i - \hat{y}_i)^2$
    - y_i — 观测值
    - $\hat{y}_i$ — 预测值
  - 结果 — 估计的斜率和截距
    - $b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$
    - $b_0 = \bar{y} - b_1\bar{x}$

2 判定系数

- 目的
  - 用来度量「估计的回归方程」的「拟合优度」
  - 应变量 y 的「变异性」能被「估计的回归方程」解释的部分 — 所占的比例
- 概念
  - 残差 — 「观测值」和「预测值」之间的「离差」 — $y_i - \hat{y}_i$
  - 误差平方和 SSE — $SSE = \Sigma(y_i - \hat{y}_i)^2$ — SST 中「未被解释」的部分
  - 总的平方和 SST — $SST = \Sigma(y_i - \bar{y})^2$
  - 回归平方和 SSR — $SSR = \Sigma(\hat{y}_i - \bar{y})^2$ — SST 中被解释的部分
  - $SST = SSR + SSE$
    - 例子
      - SST — 度量 — 观测值在直线 $\bar{y}$ 周围的聚集程度
      - SSE — 度量 — 观测值在回归线 — 周围的聚集程度
- ★ 判定系数
  - 计算公式 — $r^2 = \frac{SSR}{SST}$
    - 推理
      - 完全拟合 → SSE=0 → 比值=1
      - r^2 越接近 1，拟合程度越好
    - 内涵
      - 理解为「总的平方和」中能被「估计的回归方程」解释的百分比
      - r^2 用「百分数」表示
- 相关系数 — $r_{xy} = (\text{sign of } b_1)\sqrt{\text{Coefficient of determination}} = (\text{sign of } b_1)\sqrt{r^2}$
- 判定系数 vs 相关系数
  - 判定系数
    - 0 - 1 — 拟合优度
    - 适用更广
      - 两个 or 多个变量
      - 线性 or 非线性关系
    - 均适用
  - 相关系数
    - -1 - +1 — 线性相关强度
    - 只适用于
      - 两个变量
      - 线性关系

3 模型的假定

- 为什么要？
  - r^2 较大
    - 不能得出 — 仅能得出 — 拟合较好
    - 不能得出 — x，y 之间的关系是否显著
  - 应用「估计回归方程」前
    - 先要 — 判断「假定模型」的合理性
    - 即 — 检验「变量之间关系」的显著性
  - ★ 首先需要满足 — 「误差项」的假定
- ★「误差项」的假定
  - 关于回归模型 — $y = \beta_0 + \beta_1 x + \epsilon$ — ϵ 是「随机变量」
  - ϵ 的假定
    - $E(\epsilon) = 0$
    - 1 平均值 / 期望值为 0
      - 表明
        - β_0 and β_1 都是常数 — $E(\beta_0) = \beta_0$ and $E(\beta_1) = \beta_1$
        - 对于给定的 x 值，y 的期望值 — $E(y) = \beta_0 + \beta_1 x$
    - 2 对于所有的 x 值 — ϵ 的方差均相同
      - 记为 σ^2
      - 表明
        - y 关于「回归直线」的方差等于 σ^2
        - 对于所有 x 值 — y 的方差均相等
    - 3 — ϵ 的值是「相互独立」的 — 表明 — 对于一个「特定」的 x 值
      - 对应的「ε 值」与「其他 x 值」对应的 ε 值「不相关」
      - 对应的「y 值」与「其他 x 值」对应的 y 值「不相关」
    - 4 — ϵ 是一个「正态分布」的随机变量 — 表明 — 因为 y 是 ε 的一个线性函数
      - y 也是一个「正态分布」的随机变量
  - 总结（结合下图）
    - 1 E(y) 的值 — 根据 x 值的变化而变化
    - 2 不管 x 如何取值 — ε 和 y 的概率分布
      - 正态分布
      - 相同的方差
      - 分布形状均相同
    - 3 任一特定点上 — ε 的具体数值 — 依赖于 — 「y 的实际值」是大于还是小于 E(y)
    - ★ 注意
      - 当假定「x 和 y 之间关系」的主要部分 — $\beta_0 + \beta_1 x$
      - 同时不能忽略 — 其他形式的模型 — $y = \beta_0 + \beta_1 x^2 + \epsilon$
        - 可能是「更好」的模型
  - 回归模型的假定——图形说明



4 显著性检验

t Test

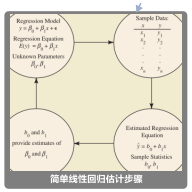
- 目的
  - 是否为 0 — 判定 β1 的值
  - 一个显著的回归关系 — 是否存在 — 检验两变量之间
- σ^2 的估计 — 需要先知道 ε 的方差
  - σ^2 也是 — 推理 — 残差 → SSE → MSE — y 关于回归直线的方差
  - 均方误差 — σ^2 的估计量
    - $s^2 = MSE = \frac{SSE}{n-2}$
    - ★ n - 2 — β0 和 β1 — 同时估计了两个参数
    - 无偏估计量
  - 估计的标准误差 — σ 的估计量
    - $s = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}}$
- 假设检验基础
  - 建立假设
    - $H_0: \beta_1 = 0$
    - $H_a: \beta_1 \neq 0$
  - β_1 的「最小二乘估计量 b1」的「抽样分布」性质
    - b_1 是 — β_1 的无偏估计量
    - 期望值 — $E(b_1) = \beta_1$
    - 标准差 — $\sigma_{b_1} = \frac{\sigma}{\sqrt{\Sigma(x_i - \bar{x})^2}}$
    - b_1 的抽样分布
    - 正态分布 — 分布形式
  - 检验统计量
    - $s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}}$ — 估计的标准差 — b_1
    - $t = \frac{b_1}{s_{b_1}}$
    - $n - 2$ — 自由度
    - $\beta_1 = 0$ — 若原假设成立
      - $t = b_1/s_{b_1}$ — 则
      - $\frac{b_1 - \beta_1}{s_{b_1}}$ — 检验统计量 — 推理
  - 拒绝法则
    - Reject H_0 if p-value $\leq \alpha$
    - Reject H_0 if $t \leq -t_{\alpha/2}$ or if $t \geq t_{\alpha/2}$
  - β_1 的置信区间
    - 点估计量 +- 边际误差 — $b_1 \pm t_{\alpha/2} s_{b_1}$
    - 拒绝 H0 — 若区间不包含 0

F Test

- 对于回归方程
  - 是否存在显著关系 — 两变量之间
  - 结论相同 — 与 t 检验
  - 仅有一个自变量
  - 「回归方程总体」的显著关系 — 只能检验 — ★ 有多个自变量
- 基于建立 — σ^2 的两个「独立」的估计量
  - MSE 均方误差
    - ★ 参数的样本估计量的期望值 等于参数的真实值 — 无偏估计量
  - MSR 均方回归
    - MSR = SSR / 回归自由度
    - 回归自由度 = 自变量个数
- 建立假设
  - $H_0: \beta_1 = 0$
  - $H_a: \beta_1 \neq 0$
- 检验统计量
  - $F = \frac{MSR}{MSE}$
    - 1 — 分子自由度
    - n - 2 — 分母自由度
  - 若原假设
    - $\beta_1 = 0$ — 成立
      - 为另一个独立估计量 — MSR
      - 接近 1 — MSR / MSE
      - 仍为无偏估计量 — MSE
    - $\beta_1 \neq 0$ — 不成立
      - 高估 σ^2 — MSR
      - 无穷大 — MSR / MSE
- 拒绝法则
  - Reject H_0 if p-value $\leq \alpha$
  - Reject H_0 if $F \geq F_\alpha$
- 简单线性回归 ANOVA 表

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	p-value
Regression	SSR	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$	
Error	SSE	$n - 2$	$MSE = \frac{SSE}{n-2}$		
Total	SST	$n - 1$			

- ★ 注意
  - 拒绝 H0
    - 两变量有「显著关系」
    - 只能说明 — x 和 y 是相关的 — 在 x 的样本观测值范围内
    - 不能说明
      - ~~有「因果关系」~~
      - ~~有「线性关系」~~ — x 的「样本观测值范围」里 — 线性关系只限于
  - 统计检验基于
    - t 检验 — b_1 抽样分布的性质
    - F 检验 — 来得到 — 误差项的假定条件
  - 统计显著性 vs 实际显著性
    - 大样本容量时，b1 值即使很小 — 也能得到 — 显著的「统计显著性」
    - 而很难说 — 多大的「实际显著性」
  - 「线性关系」的显著性检验
    - 利用
      - 样本相关系数 r_{xy}
      - 总体相关系数 ρ_{xy}
    - 假设
      - $H_0: \rho_{xy} = 0$
      - $H_a: \rho_{xy} \neq 0$
    - 结果 — 与 t 检验和 F 检验相同
    - 相关系数检验 — 不必要再利用 — 已经进行 t / F 检验



简单线性回归估计步骤