应用商务统计分析

@NWU

强喆

zhe.qng@nwu.edu.cn

2019 年 3 月 25 日

目录

1 第一章线性回归 2

1.1 第一节案例介绍 2
1.2 第二节模型定义 4
1.3 第三节描述性分析 5
1.4 第四节参数估计 6
1.5 假设检验/显著性检验 10
1.6 模型诊断 13
1.7 变量选择 17
1.8 模型预测 18

2 方差分析,Analysis of Variance, ANOVA 20

2.1 案例介绍 20
2.2 描述性分析 21
2.3 方差检验概述 22
2.4 单因素方差分析,one-way ANOVA 27
2.5 多重比较 29

课程内容

- 教程：应用商务统计分析王汉生北京大学出版社
- 成绩：
  - 平时 20%: 作业和 (1 至 2 次) 大报告
  - 期中 20%: 闭卷
  - 期末 60%: 闭卷
- 上机课：
  - 软件 Rstudio(网址：hansheng.gsm.pku.edu.cn)
  - 时间周三 3-4 节

1 第一章线性回归

1.1 第一节案例介绍

背景

- 目前中国的资本市场逐渐成熟，投资于股市成为众多企业乃至个人的重要理财方式。因此利用上市公司当年的公开的财务指标对其来年盈利状况予以预测就成为投资人最重要的决策依据。
- 本案例随机抽取深市和沪市 2002 年和 2003 年各 500 个样本，对上市公司的净资产收益率（return on equity, ROE）进行预测。

目标与变量

- 目标：盈利预测
- 因变量：下一年的净资产收益率（ROE）
- 自变量：当年的财务信息
  - ROEt: 当年净资产收益率
  - ATO: 资产周转率（asset turnover ratio）
  - LEV: 债务资本比率（debt to asset ratio）反映公司基本债务状况
  - PB: 市倍率（price to book ratio）反映公司预期未来成长率
  - ARR: 应收账款/主营业务收入（account receivable over total income）反映公司的收入质量
  - PM: 主营业务利润/主营业务收入（profit margin）反映公司利润状况
  - GROWTH: 主营业务增长率（sales growth rate）反映公司已实现的当年增长率
  - INV: 存货/资产总计（inventory to asset ratio）反映公司的存货状况
  - ASSET: （对数）资产总计（log-transformed asset）反映公司的规模
- 样本容量：2002 年 500；2003 年 500

对模型进行进一步分析

- 哪个自变量在预测方面最有用？
- 哪个自变量是最重要的？
- 如何使用模型进行预测

1.2 第二节模型定义

模型建立

- 线性回归模型：
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$
- β_j 反映了 x_{ij} 每变化一个单元对 y_i 的影响；
- ε_i 反映了模型对因变量的影响，直接反应了自变量 x_i 对因变量的预测能力。
- 模型假设：
  - 独立性假设。
    * 不同公司 X_i 之间相互独立 (一定程度上).
      反例：
      · 同一年份，不同股票价格受当前宏观经济影响
      · 母子公司之前有关联
      · 多个观测来自同一公司
    * 残差项 ε_i 与解释性变量 X_i 相互独立
  - 常方差假设。
    假设 $Var(\varepsilon_i) = c$，与当前的公司财务信息无关。公司的盈利状况波动不依赖所考虑的财务指标。
    反例，大公司方差小，高速成长的公司方差大；中心极限定理保证了，只要样本足够多，这一假设下的统计推断仍然有效；
  - 正态性假设。
    假设 ε_i 服从正态分布。
    反例，一般金融数据是重尾的 (heavy tailed). 产生极值的概率大于正态分布的预测。

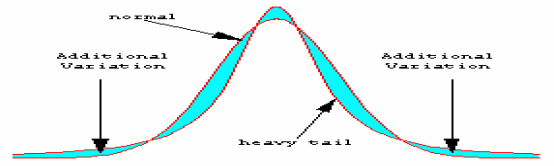


图 1: heavy tailed vs normal distribution

1.3 第三节描述性分析

- 指标：均值 Mean, 最小值 Min, 最大值 Max, 中位数 Median, 标准差 SD
- 相关性分析：(cor)
- 画出 ROEt 对 ROE 的散点图

1.4 第四节参数估计

- 模型：

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

- 最小二乘估计量/残差平方和 (Residual sum of squares,RSS)/误差平方和 (error sum of squares, SSE)

$$RSS = \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})$$

- 最小二乘估计 $\hat{\beta}$:

$$\hat{\beta} = \arg\min_{\beta} RSS$$

最小二乘估计

- 变量的矩阵形式：

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad Y = \begin{pmatrix} y_{11} \\ \cdots \\ y_{n1} \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

- 模型的矩阵形式

$$Y = X\beta + \varepsilon$$

- β 的最小二乘估计

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

- 拟合 Y 的估计

$$\hat{Y} = X\hat{\beta} = \underbrace{X(X^T X)^{-1} X^T}_{H} Y$$

  称 H 为帽子矩阵 (hat matrix)

- 残差 e 为

$$e = Y - \hat{Y} = (I - H)Y$$

$\hat{\beta}$ 的期望和方差, 抽样分布

- 期望：

$$E[\hat{\beta}] = (X^T X)^{-1} X^T E[Y] = (X^T X)^{-1} X^T X\beta = \beta$$

  因而 $\hat{\beta}$ 是 β 的无偏估计。

- 方差：

$$\begin{aligned} Var[\hat{\beta}] &= (X^TX)^{-1}X^TVar[Y]((X^TX)^{-1}X^T)^T \\ &= (X^TX)^{-1}X^TVar[Y]X(X^TX)^{-1} = (X^TX)^{-1}X^TVar[\varepsilon]X(X^TX)^{-1} \end{aligned} \tag{1.1}$$

  – 假设已知 ε 的方差为 σ^2：

$$Var[\hat{\beta}] = (X^TX)^{-1}X^T\sigma^2X(X^TX)^{-1} = \sigma^2(X^TX)^{-1} \tag{1.2}$$

  – σ^2 未知:
  由于 RSS 的期望是 $\sigma^2(n-p-1)$, 所以 $\hat{\sigma}^2$ 的无偏估计是:

$$\hat{\sigma}^2 = \frac{RSS}{n-p-1}$$

- 抽样分布：由于 X 是固定的值，Y 是随机变量 (由于正态性假设，假设 ε 服从正态分布，因而 Y 也服从正态分布，从而 X 也服从正态分布。) 这样 $\hat{\beta}$ 的抽样分布为:

$$\hat{\beta} \sim \mathcal{N}(\beta, (X^TX)^{-1}\sigma^2)$$

RSS 的期望和分布

- RSS 的期望：

$$\begin{aligned} &E[RSS] = E[\hat{e}^T\hat{e}] \\ &RSS = \hat{e}^T\hat{e} = Y^T(I-H)(I-H)Y \\ &(I-H)Y = (I-H)(X\beta+e) = X\beta+e-HX\beta-He \\ &HX = X(X^TX)^{-1}X^TX = X \end{aligned} \tag{1.3}$$

  所以有

$$\begin{aligned} &(I-H)Y = e-He = (I-H)e \\ &RSS = e^T(I-H)^T(I-H)e = e^T(I-H)e \\ &E[RSS] = E[e^T(I-H)e] = E[\sum_{i,j}(I-H)_{ij}e_ie_j] = \sum_{ij}(I-H)_{ij}E[e_ie_j] \end{aligned} \tag{1.4}$$

由于 $E[e_ie_j]$ 为零，如果 $i \neq j$; 等于 σ^2, 如果 $i = j$. 因而

$$E[RSS] = \sum_i (I-H)_{ii}\sigma^2 = \sigma^2(n - \sum_i H_{ii}) = \sigma^2(n - tr(H))$$

根据等式 $tr(AB) = tr(BA)$，因而

$$tr(H) = tr(X(X^TX)^{-1}X^T) = tr((X^TX)^{-1}X^TX) = tr(I_{p+1}) = p+1$$

所以

$$E[RSS] = \sigma^2(n-p-1)$$

- RSS 的分布：由于 $RSS = e^T(I-H)e$，其中 e 是随机变量，$e \sim \mathcal{N}(0, \sigma^2 I)$，$I-H$ 是对称矩阵，$rank(I-H) = n-p-1$, 因而

$$\frac{RSS}{\sigma^2} = \frac{e^T}{\sigma}(I-H)\frac{e}{\sigma} \sim \chi^2(n-p-1)$$

R^2——拟合优度 goodness-of-fit

- 总平方和 (Toal sum of squares,SST)
  假设要预测 y_i 而没有关于解释性变量 x_i 的任何数据，也就是只允许用 $y_1, \cdots, y_n$ 来预测 y_i。在这种情况下，显然我们要考虑使用 $\overline{y}$ 进行预测。则用这种方法预测第 i 个因变量的误差是 $y_i - \overline{y}$，而全部的误差是：

$$SST = \sum_i (y_i - \overline{y})^2$$

- 残差平方和 (SSE)
  另一方面, 如果允许使用解释性变量的数据，则预测由线性预测模型给出：

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

  第 i 个因变量的预测误差是残差 $\hat{e}_i$，全部的误差是残差平方和：

$$SSE = \sum_i (y_i - \hat{y}_i)^2$$

- 由于使用解释变量的情况总是好过不使用它们，因而 SSE 总是小于等于 SST。

- R^2:
  如果 SSE 相比 SST 非常小，以为着解释性变量在预测因变量时非常有用；另一方面，如果 SSE 只比 SST 小一点，说明通过使用解释性变量并没有真正获取多少有用的信息。R^2 量化了使用解释性变量在预测因变量时的有用程度，总是在 [0,1] 之内：

$$R^2 = 1 - \frac{SSE}{SST}$$

- 如果 R^2 很大，说明 SSE 比 SST 小很多，因而解释变量在预测方面非常有用；
- 如果 R^2 很小，说明 SSE 只比 SST 小一点，因而解释性变量在预测方面用处不大。
- 当使用的模型参数更多时，R^2 会看起来更好，但着意味着过拟合。

1.5 假设检验/显著性检验

本节有两个检验

1. 是否至少有一个财务指标对公司下一年的盈利状况有预测能力：
   H_0:$\beta_j = 0,\ \forall j = 1, \cdots, p$
   H_1: 存在至少一个 j 使得 $\beta_j \neq 0$

2. 对一个给定的自变量 x_ij, 到底那几个财务指标重要？需要逐一进行检验:
   H_0:$\beta_j = 0$
   H_1:$\beta_j \neq 0$

F 分布

$$F = \frac{\chi^2(df_1)/df_1}{\chi^2(df_2)/df_2} \sim F(df_1, df_2)$$

F 分布的随机变量是两个开方分布随机变量分别除以各自的自由度之后的比值。

一般的 F 检验

F 检验一般用于检验两个分布的方差是否相等。设第一个分布的方差为 $\hat{\sigma}_1^2$, 第二个分布的方差为 $\hat{\sigma}_2^2$，检验 $F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} ? = 1$

用于线性回归模型的 F 检验

- 对于线性回归，
$$\frac{SSE}{\sigma^2} \sim \chi^2(n - p - 1)$$
- 则根据 F 分布的定义，F 统计量应为
$$F = \frac{SSE_1/(\sigma_1^2 df_1)}{SSE_2/(\sigma_2^2 df_2)}$$
这是课本中的定义, 其中 df_1, df_2 分别是 SSE_1, SSE_2 的自由度。
- 又根据
$$\hat{\sigma}^2 = \frac{SSE}{df}$$
所以 F 统计量又可以表示为
$$F = \frac{\hat{\sigma}_1^2/\sigma_1^2}{\hat{\sigma}_2^2/\sigma_2^2}$$
在原假设 H_0:$\sigma_1^2 = \sigma_2^2$，假设所比较的两个分布方差相等时，F 统计量又表示为
$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$$
这是 F 统计量一般的定义。

F 检验用于本节第 1 个检验

- 全模型 M:$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$
  去掉所有参数 β_j(所有的 $\beta_j = 0$) 的模型 m:$y_i = c + \varepsilon_i$
- 这样 $RSS(m) \geq RSS(M)$，因为模型 M 更加精细，预测误差更小。
- F 统计量为
$$F = \frac{(RSS(m) - RSS(M))/p}{RSS(M)/(n-p-1)}$$
因为 $RSS(m) - RSS(M) \sim \chi^2(p)$
- 实际上，$RSS(M) = SSE$, 而对模型 m，显然截距项 c 的最大似然估计或者最小二乘估计应是均值 $\overline{y}$，因而
$$RSS(m) = \sum_{i=1}^{n}(y_i - \hat{c})^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2 = SST$$
这样 $F = \frac{(SST-SSE)/p}{SSE/(n-p-1)} \sim F(p, n-p-1)$，与书上的表达一致。

F 检验用于本节第２个检验

- 全模型 M:$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$
  去掉第 j 个参数 $\beta_j(\beta_j = 0)$ 的模型 m:$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{j-1} x_{i,j-1} + \beta_{j+1} x_{i,j+1} + \cdots + \beta_p x_{ip} + \varepsilon_i$
- 这样 $RSS(m) \geq RSS(M)$，因为模型 M 更加精细，预测误差更小。
- F 统计量为
$$F = \frac{(RSS(m) - RSS(M))/1}{RSS(M)/(n-p-1)} \sim F(1, n-p-1)$$
因为 $RSS(m) - RSS(M) \sim \chi^2(1)$。

作业

- 用 F 检验做本节第 2 个假设检验
- 用 R 语言或者其他软件 (Excel) 等编程实现，打印报告。

t 检验用于本节第 2 个检验

- 由于 $\hat{\beta} \sim \mathcal{N}_{p+1}(\beta, \sigma^2(X^TX)^{-1})$，也就是说 $\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 v_j(X))$, $v_j(X)$ 是矩阵 $(X^TX)^{-1}$ 的第 j 个对角元。
- 原假设 $H_0 : \beta_j = 0$,　这样 $\frac{\hat{\beta}_j}{\sigma\sqrt{v_j(X)}} \sim \mathcal{N}(0,1)$;

- 要用样本构造一个统计量 T, 但 σ 未知，用 $\hat{\sigma}$ 代替 σ, 则统计量 $T = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j(X)}}$;
- 分子分母同时除以 $\sigma\sqrt{v_j(X)}$:

$$T = \frac{\hat{\beta}_j/(\sqrt{v_j(X)}\sigma)}{\hat{\sigma}/\sigma} = \frac{\hat{\beta}_j/(\sqrt{v_j(X)}\sigma)}{\sqrt{RSS/(n-p-1)/\sigma^2}}$$

这样分子 $\frac{\hat{\beta}_j}{\sqrt{v_j(X)}\sigma} \sim \mathcal{N}(0,1)$, 分母 $\sqrt{\chi^2(n-p-1)/(n-p-1)}$，因而统计量 $T \sim t(n-p-1)$。

1.6 模型诊断

本节模型诊断包括四个模块检验：

- 异方差检验【通过画出残差 vs 拟合值的图】
- 残差正态性检验【QQ 图】
- 异常值检验【Cook 距离】
- 多重共线性程序【VIF 方差膨胀因子】

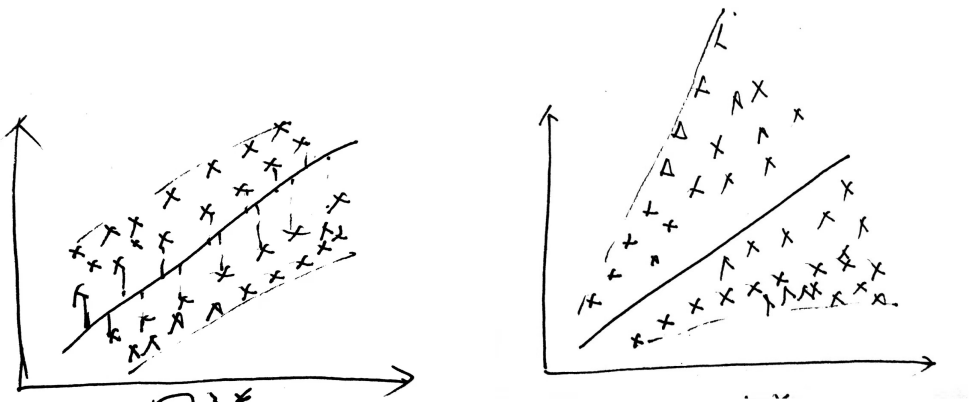
实际上回忆第一节的模型假设

- 独立性
- 常方差性
- 正态性

异方差检验就是检验常方差性：是否所有公司的噪声 ε_i 的方差相同？即 $var(\varepsilon_i) = ?\sigma^2$;
而残差正态性检验是为了检验残差的正态性：是否 ε_i 服从正态分布？

异方差检验

同方差异方差演示



(a) 同方差，各个 X_i 方差在一条宽度相同的带子里

(b) 异方差，各个 X_i 的方差越来越大，这里 $var(\varepsilon_i) = f(X_i)$ 关于 X_i 是增函数。

图 2: 同方差 vs 异方差

残差正态性检验

QQ 图:把样本残差 ε_i 按照从大到小的顺序排序,以 $\varepsilon_{(1)}, \varepsilon_{(2)}, \cdots, \varepsilon_{(n)}$ 为 X 轴，以 $z_{1/n}, z_{2/n}, \cdots, z_{n/n}$ 为

异常值检验

- 一般而言 cook 距离定义如下:

$$D_i = \frac{\sum_{i=1}^{n}(\hat{y}_j - \hat{y}_j^{(i)})^2}{\hat{\sigma}^2 \cdot p}$$

  其中，$\hat{y}_j$ 是用全模型拟合的第 j 个点的值，$\hat{y}_j^{(i)}$) 是用去掉第 i 个点拟合的第 j 个点的值，p 为模型的参数个数。

- 这里由于是线性模型，$\hat{y} = X\hat{\beta}$, $\hat{y}^{(i)} = X\hat{\beta}^{(i)}$, 则 cook 距离变成

$$\begin{aligned} D_i &= \frac{(\hat{y} - \hat{y}^{(i)})^T(\hat{y} - \hat{y}^{(i)})}{\hat{\sigma}^2 \cdot (p+1)} \\ &= \frac{(X\hat{\beta} - X\hat{\beta}^{(i)})^T(X\hat{\beta} - X\hat{\beta}^{(i)})}{\hat{\sigma}^2 \cdot (p+1)} \\ &= \frac{(\hat{\beta} - \hat{\beta}^{(i)})^T X^T X(\hat{\beta} - \hat{\beta}^{(i)})}{\hat{\sigma}^2 \cdot (p+1)} \end{aligned}$$

- Cook 距离反应了第 i 个观测对整个估计的影响力：如果删去第 i 个点与全模型相比，拟合的残差平方和变化不大，说明第 i 个点不能对模型造成影响，因而不是异常值；如果删去第 i 个点与全模型相比，拟合的残差平方和变化很大，说明第 i 个点对模型影响较大，考虑它是异常值。

- Cook 距离只是给出一个参考指标，并不是严格的指标，具体要看要分析的问题。

多重共线性

- model:$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$,

- 考虑其中一个变量是否能被其余的几个变量线性组合表示。如果可以表示，那么在线性回归中，这个系数的作用是双倍的。

- 考虑模型 $m_j : x_{ij} = \delta_0 + \delta_1 x_{i1} + \cdots + \delta_{i,j-1} x_{i,j-1} + \delta_{i,j} x_{i,j+1} + \delta_{i,p-1} x_{ip} + \nu_i$ 是用其他变量线性预测第 j 个变量。

- 利用最小二乘，计算出 $\hat{x}_{ij}$

- 定义
$$R_j^2 = 1 - \frac{\sum_{i=1}^n (x_{ij} - \hat{x}_{ij})^2}{\sum_{i=1}^n (x_{ij} - \overline{x}_{ij})^2} = 1 - \frac{RSS_j}{TSS_j}$$
是第 j 个变量的 R^2;

- 定义 VIF(方差膨胀因子):
$$VIF = \frac{1}{1 - R_j^2}$$
这样 RSS_j 越小，R_j^2 越大，VIF 越大，说明模型 m_j 越有用，则第 j 个变量可以被其余的几个变量线性表示，说明共线性性很强，可以删去这个变量。

推导 VIF

- 计算 $var(\hat{\beta}_j) = \sigma^2 v_j(X)$, $v_j(X)$ 表示 $(X^TX)^{-1}$ 的第 j 个对角元
- 根据 schur-complement 公式:
$$\begin{pmatrix} A & B \\ B^T & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}B^T)^{-1} & * \\ * & (D - B^TA^{-1}B)^{-1} \end{pmatrix}$$
- X_j 表示 X 的第 j 列，则
$$X^TX = \begin{pmatrix} X_j^TX_j & X_j^TX_{-j} \\ X_{-j}^TX_j & X_{-j}^TX_{-j}^T \end{pmatrix}$$
-
$$(X^TX)^{-1} = \begin{pmatrix} [X_j^TX_j - (X_j^TX_{-j}) \cdot (X_{-j}^TX_{-j})^{-1} \cdot (X_{-j}^TX_j)]^{-1} & * \\ * & * \end{pmatrix}$$
因而 $v_j(X) = [X_j^TX_j - (X_j^TX_{-j}) \cdot (X_{-j}^TX_{-j})^{-1} \cdot (X_{-j}^TX_j)]^{-1}$
- 对模型 m_j 进行最小二乘估计:
$$X_j = X_{-j}\delta + \nu$$
$$\hat{\delta} = \underbrace{(X_{-j}^TX_{-j})^{-1}X_{-j}^TX_j}_{(X^TX)^{-1}X^TY}$$
$$\hat{X}_j = X_{-j}\hat{\delta} = \underbrace{X_{-j}(X_{-j}^TX_{-j})^{-1}X_{-j}^T}_{H}X_j$$
- 这样
$$\begin{aligned} v_j(X) &= \underbrace{[X_j^TX_j - (X_j^T\underbrace{X_{-j}}_{H}) \cdot (X_{-j}^TX_{-j})^{-1} \cdot (X_{-j}^T}X_j)]^{-1}}_{Y^TY - Y^THY} \\ &= \frac{1}{\hat{\nu}^T\hat{\nu}} = \frac{1}{RSS_j} \end{aligned}$$

$$
\begin{aligned}
var(\hat{\beta}_j) &= \frac{\sigma^2}{RSS_j} \\
&= \frac{\sigma^2}{TSS_j}\frac{TSS_j}{RSS_j} \\
&= \frac{\sigma^2}{\sum_{i=1}^{n}(x_{ij}-\overline{x}_{ij})^2}\underbrace{\frac{1}{1-R_j^2}}_{VIF}
\end{aligned}
$$

- VIF 与 R_j^2 成正比。R_j^2 越大，$\hat{\beta}_j$ 的方差越大，说明 RSS_j 与 TSS_j 的差距越大，X_j 的可解释性越强。

1.7 变量选择

- AIC:
$$AIC = n\{\log\left(\frac{RSS}{n}\right) + 1 + \log(2\pi)\} + 2 \times (p+1)$$
- BIC:
$$BIC = n\{\log\left(\frac{RSS}{n}\right) + 1 + \log(2\pi)\} + \log(n) \times (p+1)$$
- 当 $n \geq 8$ 时，$\log(n) > 2$,BIC 的惩罚项比 AIC 惩罚力度大，因此 AIC 选出的模型变量个数多于 BIC 选出的个数。

1.8 模型预测

- model:$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$
- 实际上，置信区间 = 样本估计值 ±t 分位数 × 标准差，因而分别计算样本估计值和方差即可计算预测区间。
- 对于新的观测点 x_0, 点估计是

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \cdots + \hat{\beta}_p x_{0p} = X_0 \hat{\beta}$$

  (但实际上满足线性模型 $\hat{y}_0 = X_0\hat{\beta} + \varepsilon$)

- $\hat{y}_0$ 的方差：

$$\begin{aligned} var(\hat{y}_0) = var(X_0\hat{\beta}) + var(\varepsilon) &= X_0^T var(\hat{\beta}) X_0 + \sigma^2 \\ &= X_0^T (X^T X)^{-1} \sigma^2 X_0 \ + \sigma^2 = \sigma^2 (1 + X_0^T (X^T X)^{-1} X_0 \ ) \end{aligned} \quad (1.5)$$

  这是因为 $\hat{\beta} \sim \mathcal{N}(\beta, (X^T X)^{-1}\sigma^2)$.

- 则 $\hat{y}_0$ 的预测区间是：

$$X_0\hat{\beta} \pm \text{t 分位数} \times \sigma\sqrt{1 + X_0^T (X^T X)^{-1} X_0}$$

- $E(\hat{y}_0)$ 的点估计:

$$E(\hat{y}_0) = X_0 E(\hat{\beta}) = X_0\beta$$

  则 $E(\hat{y}_0)$ 的点估计是 $X_0\hat{\beta}$

- $E(\hat{y}_0)$ 的方差：

$$var(X_0\hat{\beta}) = X_0^T var(\hat{\beta}) X_0 = X_0^T (X^T X)^{-1} \sigma^2 X_0 = X_0^T (X^T X)^{-1} X_0 \sigma^2$$

- 则 $E(\hat{y}_0)$ 的预测区间是：

$$X_0\hat{\beta} \pm \text{t 分位数} \times \sigma\sqrt{X_0^T (X^T X)^{-1} X_0}$$

作业

编程计算 $\hat{y}_0$ 的预测区间和 $E(\hat{y}_0)$ 的置信区间，并画图。

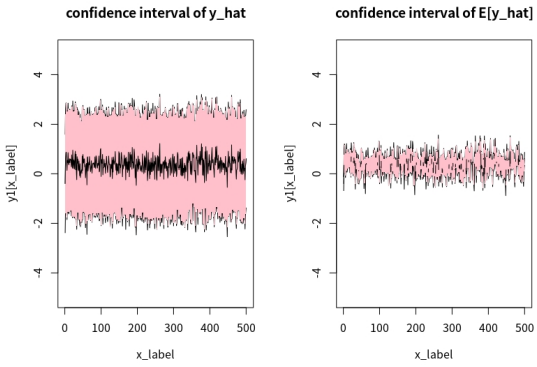

confidence interval of y_hat
confidence interval of E[y_hat]
y1[x_label]
x_label
0
100
200
300
400
500
-4
-2
0
2
4

图 3

2 方差分析,Analysis of Variance, ANOVA

2.1 案例介绍

- 回归分析：因变量是连续型因素 (如年龄、收入、价格等)
- 方差分析：因变量是离散型因素 (如性别、职业、种族等)

案例介绍

- 北京市房地产
- 从搜房网随机选取 2003-2004 年度新开盘楼盘共 506 个
- 清楚不完整或有明显错误的数据后，最终得到 200 个合格的楼盘样本

研究目标

- 房价
- 影响房价的因素
- 暂不考虑连续型变量

2.2 描述性分析

命令

- boxplot: 画出每一个水平下的中位数、样本方差，进行观察
- 如果不满足同方差性，对 price 进行对数变换,$log.price = \log(price)$, 并画图 boxplot
- summary: 查看各水平下样本的分布是否均匀

2.3 方差检验概述

- 假设要比较三个学校的学生成绩。设为 school1,school2,school3，school1 有 1 万人，school2 有 8000 人，school3 有 1 万 2000 人；
- 假设检验：
  - 原假设 H_0： $\mu_1 = \mu_2 = \cdots = \mu_K$
  - 备择假设 H_1： 至少有一个均值和其它不同
- 对每个学校全部学生进行调查文件既花费时间又花费金钱，因此考虑从每个学校的学生中选择样本。school1 选择 10 个学生，school2 选择 8 个学生，school3 选择 12 个学生。用样本均值代替总体均值。如图1

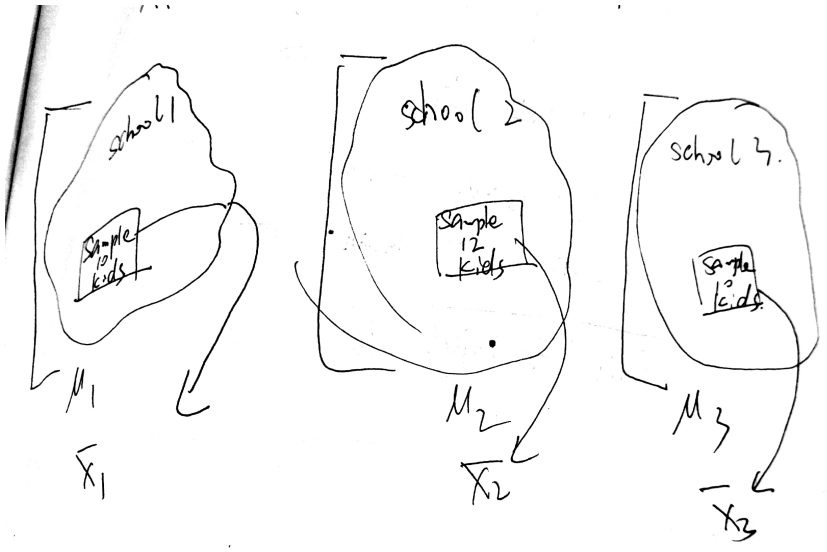


图 1: ANOVA，三个学校的学生成绩进行比较

- 画出直方图来看。
  - 第一种情况，如图2 原假设 H_0： $\mu_1 = \mu_2 = \cdots = \mu_K$ 看起来可能发生
    备择假设 H_1： 至少有一个均值和其它不同
    拒绝 H_0 失败
  - 第二种情况, 如图3：原假设 H_0： $\mu_1 = \mu_2 = \cdots = \mu_K$ 看起来不太可能发生
    备择假设 H_1： 至少有一个均值和其它不同
    拒绝 H_0

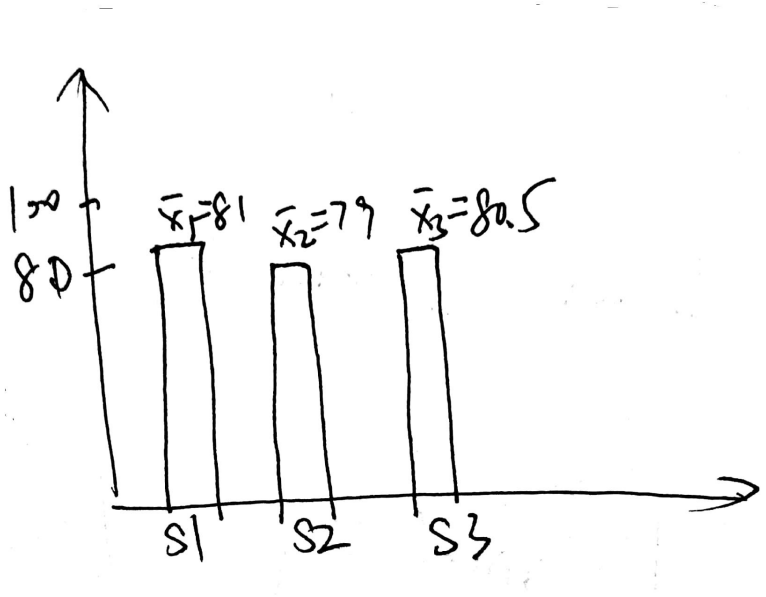


图 2: 原假设 H_0： $\mu_1 = \mu_2 = \cdots = \mu_K$ 看起来可能发生
备择假设 H_1： 至少有一个均值和其它不同
拒绝 H_0 失败

- 方差分析是研究均值的不同，均值之间有多少变化，考虑的是均值的 variability，而非研究方差，但它名字叫做方差分析、
- ANOVA 至告诉我们其中一个或者几个不同，但没有说明哪一个不同。所以要做另外的检验
- 数据的数量和质量影响结果，例如有很多异常值，必然会影响平均。例如在 case2 中，如果 school1 选择的 10 个学生中有两个成绩不好的同学没有去参加调查，成绩没有记入样本平均，school2 中有两个学生没有答卷，得 0 分，使得 school2 的样本均值偏小。

均值
均值有两种

- 总体均值

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i} x_{ij}}{\sum_{i=1}^{K} n_i}$$

K 是 population/group/treatment 的数目
n_i 是第 i 个组的样本大小

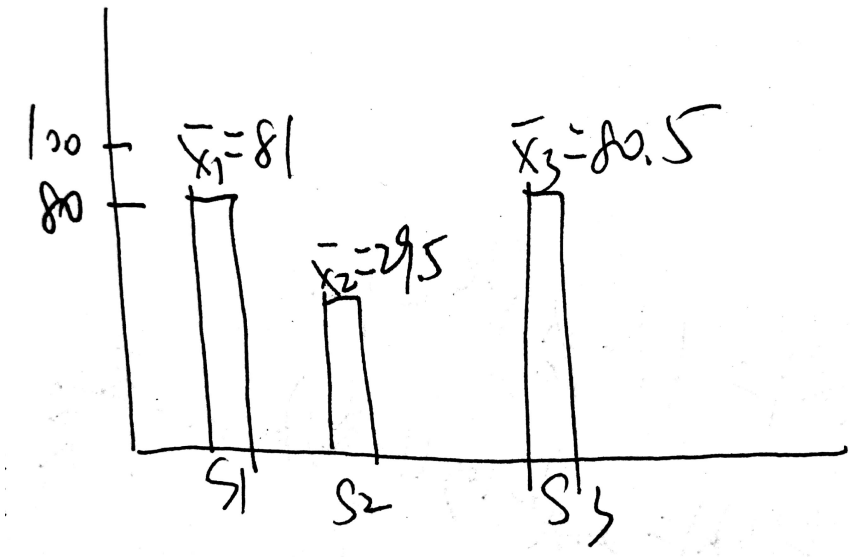


图 3: 原假设 H_0： $\mu_1 = \mu_2 = \cdots = \mu_K$ 看起来不太可能发生
备择假设 H_1： 至少有一个均值和其它不同
拒绝 H_0

- 第 i 组的样本均值

$$\overline{x}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i}$$

实际上，均值代表的是一个比较的基准线，在所选的样本中，有的比基准线大，有的比基准线小。

方差

回忆方差的估计式是

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

代表样本离基准线的离散程度。样本变化范围越大，方差越大，样本越集中，方差越小。观察这个表达式，分子是样本离基准线的误差平方和 (sum of squares, SS), 分母是分子的自由度。

这里方差也有两种

1. 组间方差 (between-variance)

$$S^2(B) = \frac{SS(B)}{d.f.(B)}$$

组间平方和 (between sum of squares,SS(B)) 是

$$SS(B) = \sum_{i=1}^{k} n_i(\overline{x}_i - \overline{\overline{x}})^2$$

总体的基准线是 $\overline{\overline{x}}$, 第 i 组的样本代表是 $\overline{x}_i$, 因而是 $(\overline{x}_i - \overline{\overline{x}})^2$, 并对 i 求和。
但每个组的样本数目不同，这里给每个组赋予的权重是每个组的样本数目 n_i, 这样就是 $n_i(\overline{x}_i - \overline{\overline{x}})^2$, $i = 1, \cdots, n_i$, 并对 i 求和。
自由度是 $d.f.(B)$, 共有 K 个 $\overline{x}_i$，一个条件 $\overline{\overline{x}}$，因而自由度是 $K-1$

2. 组内方差 (within-variance)

$$S^2(W) = \frac{SS(W)}{d.f.(W)}$$

组内平方和 (within sum of squares, SS(W)) 是

$$SS(W) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)^2$$

对第 i 个组，第 j 个样本是 x_{ij}, 第 i 个组的基准线是 $\overline{x}_i$，对所有样本求和 $i = 1, \cdots, K$, $j = 1, \cdots, n_i$
自由度是 d.f.(W)，共有 n 个样本，K 个基准线，所以是 $n - K$

F 统计量
$F = \frac{S^2(B)}{S^2(W)}$ 这是一个 F 统计量，回忆这里的原假设和备择假设：

- 原假设 H_0： $\mu_1 = \mu_2 = \cdots = \mu_K$
- 备择假设 H_1： 至少有一个均值和其它不同

ANOVA 总是右尾检验：如图4

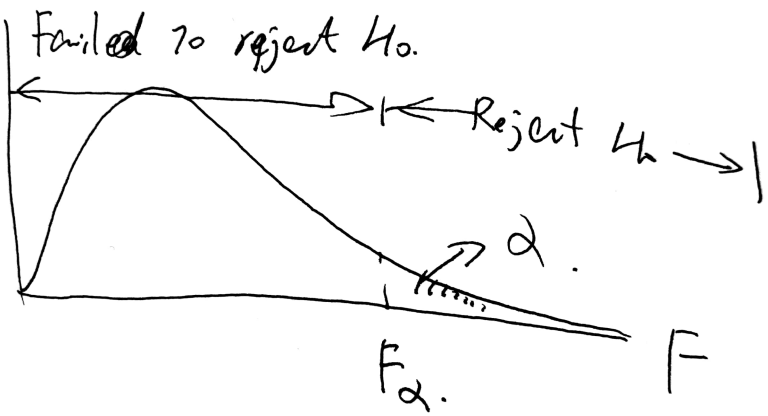


图 4: ANOVA 总是右尾检验

如果原假设成立，$\mu_1 = \mu_2 = \cdots = \mu_k$, 则 $\overline{x}_1 \approx \overline{x}_2 \approx \cdots \approx \overline{x}_k$。则组间变化几乎为零，$SS(B) \approx 0$, 因而 F 统计量非常小，落入大概率区域，也就是落入拒绝 H_0 失败的区域。拒

绝原假设失败，不能拒绝原假设，则接受原假设。
如果原假设不成立，μ_k 的不同 $\Rightarrow \overline{y}_k$ 的不同 $\Rightarrow$ F 的增加。组间变化较大，$SS(B)$ 较大，因而 F 统计量非常大，落入小概率的阴影区域，也即是落入拒绝 H_0 的区域。则拒绝原假设。

2.4 单因素方差分析,one-way ANOVA

研究目标

- 对于某一因素 (例如环线位置)，为了研究它在不同水平下因变量的均值是否相等

模型

$$y_{ij} = u_i + \varepsilon_{ij}$$

- y_{ij} 表示水平 i 下第 j 个楼盘的均价
- 水平 i 下房价的均值 $\frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$
- ε_{ij} 水平 i 下第 j 个楼盘的实际价格与均值之间的残差

模型假设

- 同方差性，假设 $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$
- 残差 ε_{ij} 服从正态分布。

F 检验

- 假设检验
  1. 原假设 H_0:$u_1 = u_2 = \cdots = u_k$
  2. 备择假设 H_1:not H_0
- 统计量 $F = \frac{\text{variance between samples}}{\text{variance within samples}}$.
  如果 $\mu_1 = \mu_2 = \cdots = \mu_k$, 则 $\overline{y}_1 \approx \overline{y}_2 \approx \cdots \approx \overline{y}_k$。
  μ_k 的不同 $\Rightarrow$ $\overline{y}_k$ 的不同 $\Rightarrow$ F 的增加。
- Between sample variability(样本之间的变化):$SS(B)$ sum of squares between groups

$$SS(B) = \sum_{i=1}^{k} n_i(\overline{y}_i - \overline{y})^2$$

- Between sample variance:($MS(B)$ 组间均方差，$S^2(B)$ 组间方差)

$$S^2(B) = MS(B) = \frac{SS(B)}{d.f.(B)} = \frac{\sum_{i=1}^{k} n_i(\overline{y}_i - \overline{y})^2}{k-1}$$

  组间变化除以自由度

- Within sample variability(样本内的变化):

$$SS(W) = \sum_{i=1}^{K} SS_i = SS_1 + SS_2 + \cdots + SS_k$$

$$SS_i = \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2$$

这样

$$SS(W) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2$$

- Within sample variance(组内方差):($MS(W)$ 组内均方差，$S^2(W)$ 组内方差)

$$S^2(W) = MS(W) = \frac{SS(W)}{d.f.(W)} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2}{n-k}$$

- 这样

$$F = \frac{S^2(B)}{S^2(W)} = \frac{\frac{\sum_{i=1}^{k} n_i (\overline{y}_i - \overline{y})^2}{k-1}}{\frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2}{n-k}}$$

- 实际上总方差 = 组内方差 + 组间方差：

$$\begin{aligned} TSS &= \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \overline{y})^2 \\ &= \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \overline{y}_i + \overline{y}_i - \overline{y})^2 \\ &= \sum_{i=1}^{k} \sum_{j=1}^{n} \left[(y_{ij} - \overline{y}_i)^2 + (\overline{y}_i - \overline{y})^2 + 2(y_{ij} - \overline{y}_i)(\overline{y}_i - \overline{y}) \right] \\ &= \underbrace{\sum_{i=1}^{k} \sum_{j=1}^{n} \left[(y_{ij} - \overline{y}_i)^2 \right]}_{S^2(W)} + \underbrace{\sum_{i=1}^{k} n_i (\overline{y}_i - \overline{y})^2}_{S^2(B)} \end{aligned}$$

- 如果原假设正确，则

$$S^2(B) \sim \chi^2(k-1), \quad S^2(W) \sim \chi^2(n-k)$$

则统计量

$$F = \frac{S^2(B)}{S^2(W)} \sim F(k-1, n-k)$$

2.5 多重比较

- 四对检验
  - 2 环以内 vs 2-3 环
  - 2 环以内 vs 3-4 环
  - 2 环以内 vs 4-5 环
  - 2 环以内 vs 5 环以外
- 如果每个检验的显著水平都是 5%, 则观察 summary(lm1) 的结果。2 环以内与 2-3 环和 3-4 环的房价没有显著性差异，而与 4-5 环和 5 环以外有显著性差异。
- 如果我们进行这四对检验，如果每个检验的显著水平是 5%, 则
  - $H_0:\ \mu_1=\mu_2, \mu_1=\mu_3, \mu_1=\mu_4, \mu_1=\mu_5$
  - H_1 : 至少有一个 $\mu_k \neq \mu_1, k=2,\cdots,5$

  犯第一类错误的概率是:

$$\begin{aligned} p(\text{type I error}) &= p(\text{reject}H_0|H_0) \\ &= 1-p(\text{failed to reject}H_0|H_0) \\ &= 1-\prod_{i=2}^{K} p(\text{failed to reject}\mu_1=\mu_i|H_0) \\ &= 1-\prod_{i=2}^{K}[1-p(\text{reject}\mu_1=\mu_i|\mu_1=\mu_i)] \\ &= 1-\prod_{i=2}^{K}(1-\alpha) = 1-(1-0.05)^4 = 18.54938\% \end{aligned}$$

- 因此，一种方法是考虑 Bonferroni 修正。令每个检验的显著水平是 α/K, 这样，四个检验犯第一类错误的概率是:

$$\begin{aligned} p(\text{type I error}) &= p(\text{reject}H_0|H_0) \\ &= 1-(1-\frac{\alpha}{K})^K \end{aligned}$$

  可以证明 Bonferroni 不等式:

$$1-(1-\frac{\alpha}{K})^K \leq \alpha$$

  证明留作作业

作业

证明 Bonferroni 不等式:

$$1-(1-\frac{\alpha}{K})^K \leq \alpha$$