应用商务统计分析

@NWU

强喆

zhe.qng@nwu.edu.cn

2019 年 6 月 17 日

目录

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。

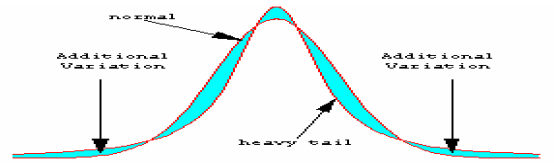


图 1: heavy tailed vs normal distribution

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】

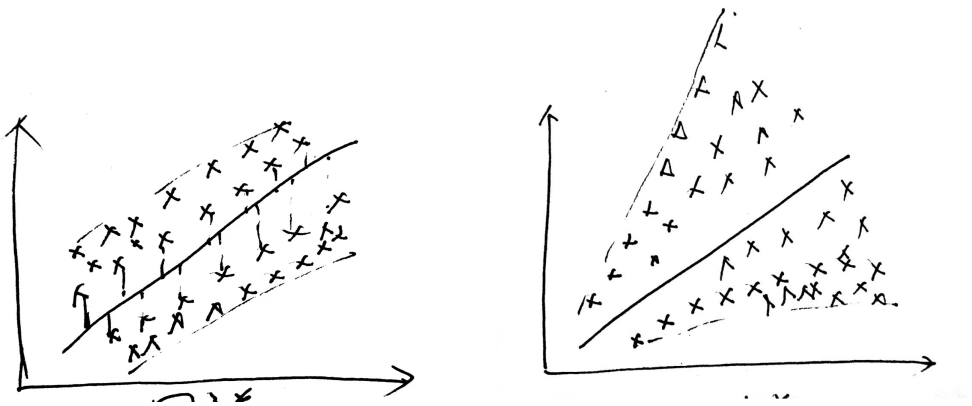
实际上回忆第一节的模型假设

- 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。

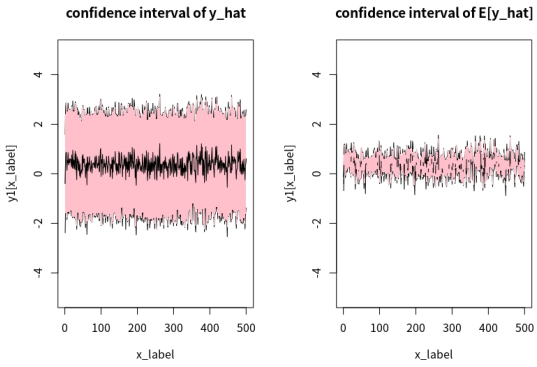

confidence interval of y_hat
confidence interval of E[y_hat]
y1[x_label]
x_label
0
100
200
300
400
500
-4
-2
0
2
4

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

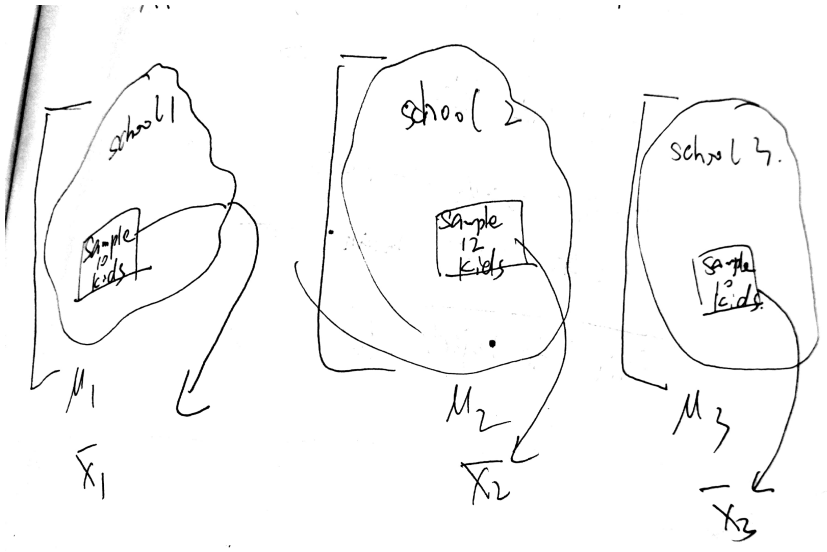


图 1: ANOVA，三个学校的学生成绩进行比较

- 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

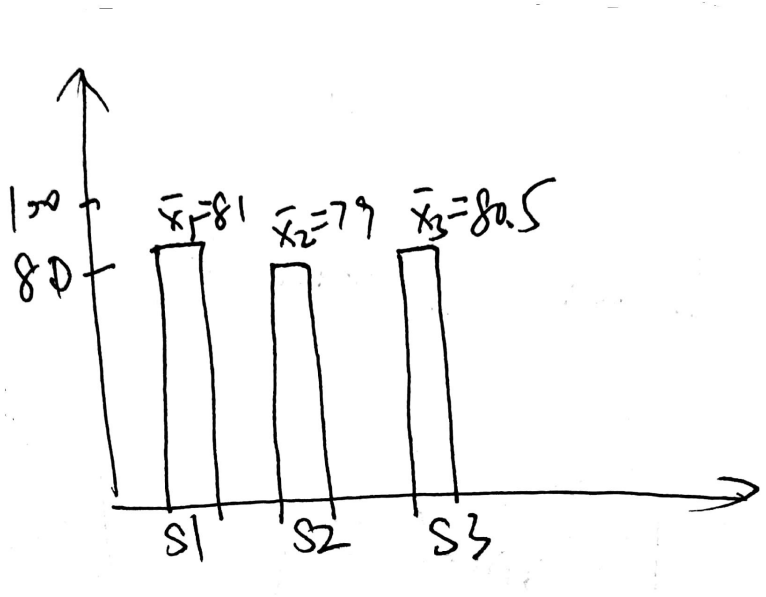


图 2: 原假设 H_0： $\mu_1 = \mu_2 = \cdots = \mu_K$ 看起来可能发生
备择假设 H_1： 至少有一个均值和其它不同
拒绝 H_0 失败

- 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

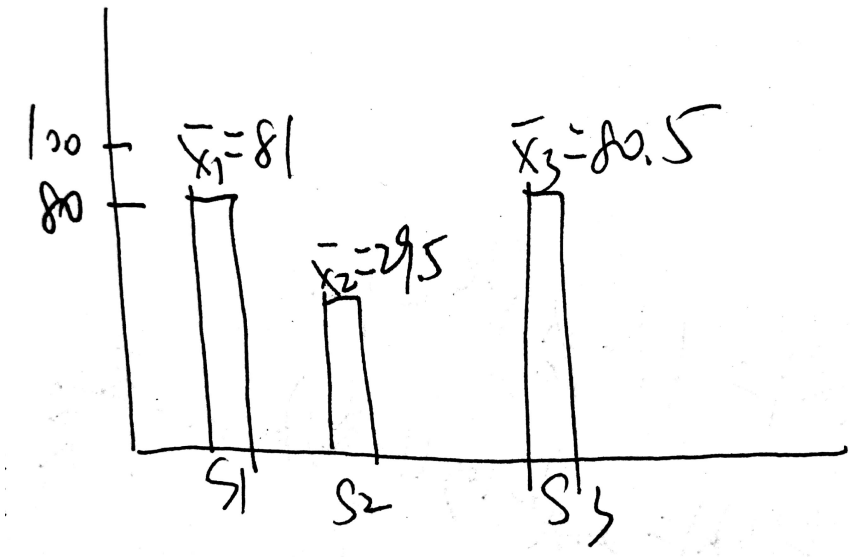


图 3: 原假设 H_0： $\mu_1 = \mu_2 = \cdots = \mu_K$ 看起来不太可能发生
备择假设 H_1： 至少有一个均值和其它不同
拒绝 H_0

- 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

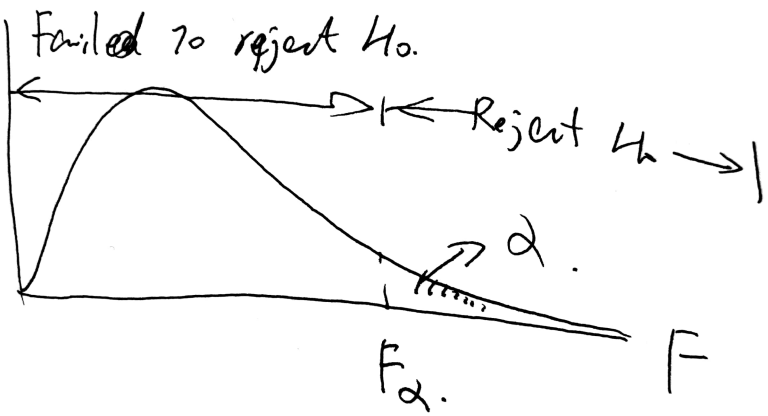


图 4: ANOVA 总是右尾检验

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:

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

2.6 双因素简单可加模型

- 单因素方差分析：考虑一个自变量的影响

```
lm1 = lm(log.price ~as.factor(ring))
```

- 双因素模型：为了考虑两个或多个自变量的共同影响，本节建立双因素简单可加模型
- 模型：

$$y_{lkr} = \mu + \alpha_l + \beta_k + e_{lkr}$$

  - 因素 A:$l = 1, \cdots, g$
  - 因素 B:$k = 1, \cdots, b$
  - 组合共:$g * b$
  - 每个组合有 n 个观测:$r = 1, \cdots, n$
  - $\mu + \alpha_l$: 因素 A 在水平 l 下的均值
  - $\mu + \beta_k$: 因素 B 在水平 k 下的均值
  - $\mu + \alpha_l + \beta_k$: 因素 A 在水平 l 下、因素 B 在水平 k 下的均值
  - $(\mu + \alpha_l + \beta_{k_1}) - (\mu + \alpha_l + \beta_{k_2}) = \beta_{k_1} - \beta_{k_2}$: 固定因素 A 水平为 l、因素 B 从 $k_1 \to k_2$

- 代码：

```
lm2.1 = lm( log.price ~ as.factor(ring) + as.factor(wuye))
Anova(lm2.1)
```

- 实验

2.7 双因素交互作用

- 模型:

$$y_{lkr} = \mu + \alpha_l + \beta_k + +\gamma_{lk} + e_{lkr}$$

  - 因素 A:$l = 1, \cdots, g$
  - 因素 B:$k = 1, \cdots, b$
  - 组合共:$g * b$
  - 每个组合有 n 个观测:$r = 1, \cdots, n$
  - $\mu + \alpha_l$: 因素 A 在水平 l 下的均值
  - $\mu + \beta_k$: 因素 B 在水平 k 下的均值
  - $\mu + \alpha_l + \beta_k + \gamma_{lk}$: 因素 A 在水平 l 下、因素 B 在水平 k 下的均值
  - $(\mu + \alpha_l + \beta_{k_1} + \gamma_{lk_1}) - (\mu + \alpha_l + \beta_{k_2} + \gamma_{lk_2}) = (\beta_{k_1} - \beta_{k_2}) + (\gamma_{lk_1} - \gamma_{lk_2})$: 固定因素 A 水平为 l、因素 B 从 $k_1 \to k_2$
    因而不能简单叠加各个因素的效应

- 代码:

```
lm2.2 = lm( log.price ~ as.factor(ring) * as.factor(wuye))
Anova(lm2.2)
```

- 实验

2.8 多因素方差分析

- 因素：城区、环线位置、物业类型、装修状况及建筑类型，如果是双因素简单可加模型，共 $C_5^2 = 10$ 种选择
- 这里考虑交互项：城区和环线
- 代码

```
lm4 = lm(log.price ~ as.factor(dis)*as.factor(ring)
 + as.factor(wuye) + as.factor(fitment) + as.factor(contype))
summary(lm4)
```

- 实验

3 协方差分析，Analysis of Covariance(ANCOVA)

- 回归分析：连续型协变量
- 方差分析：离散型协变量
- 协方差分析：离散型 + 连续型协变量

3.1 数据描述

数据描述

- 因变量：课程评估得分 (score)
- 协变量：
  - 教师职称 (title)
  - 教师性别 (gender)
  - 学生类别 (student)
  - 年份 (year)
  - 学期 (semeter)
  - 学生人数 (size)

研究目标

- 分析课程若干因素对于课程评估得分的影响

3.2 描述性分析

```
attach(a)
plot(size,score)
boxplot(score ~ ceiling(size/20))
table(ceiling(size/20))

group = 1*(size<=20)

boxplot(score~title,main="职称")
boxplot(score~gender,main="性别")
boxplot(score~student,main="学生类别")
boxplot(score~year,main="年份")
boxplot(score~semester,main="学期")
boxplot(score~group,main="班级规模")
```

3.3 单因素可加模型

- 只考虑一个离散型协变量 (group) 和一个连续性协变量 (size)
- 只考虑 (score　size)

$$score = \alpha + \beta * size + \varepsilon$$

不管大班还是小班，score 与 size 之间服从相同的等式关系，不合理

- 根据班级规模的不同，定义不同的教学评估成绩对学生人数的回归直线 (斜率 (β), 截距 (α))
- 先考虑班级规模对截距的影响：

$$score = \alpha_0 + \alpha_1 * group + \beta * size + \varepsilon$$

$$\begin{cases} \text{score} = (\alpha_0 + \alpha_1) + \beta * \text{size} + \varepsilon, & \text{if group} = 1 \\ \text{score} = \alpha_0 + \beta * \text{size} + \varepsilon, & \text{if group} = 0 \end{cases}$$

- 斜率相同，表示不管班级规模大小，学生人数单位变化所带来的教评成绩变化是一样的 (不太合理，暂作此假定)

```
lm1=lm(score~as.factor(group)+size)
summary(lm1)
```

3.4 单因素交互作用

- 不仅截距不同，斜率也不相同.

$$score = \alpha_0 + \alpha_1 * group + \beta_0 * size + \beta_1 * group * size + \varepsilon$$

$$\begin{cases} \text{score} = (\alpha_0 + \alpha_1) + (\beta_0 + \beta_1) * \text{size} + \varepsilon, & \text{if group} = 1 \\ \text{score} = \alpha_0 + \beta_0 * \text{size} + \varepsilon, & \text{if group} = 0 \end{cases}$$

-
```
  lm2=lm(score~as.factor(group)*size)
library(car)
Anova(lm2,type="III")
```

3.5 多因素协方差分析

- model: score = title + gender + student + year + semester + group * size + ε
- ε 是无法被这些客观因素解释的教评成绩，与原始成绩相比，ε 是剔除了这些客观因素影响后的教评成绩，因而更能反映教员的努力程度与授课效果。
-
```
lm3.1=lm(score~as.factor(title)+as.factor(gender)+as.factor(student)
+as.factor(year)+as.factor(semester)+as.factor(group)*size)
Anova(lm3.1,type="III")
```

- 剔除不显著的因素

```
lm3.2=lm(score~as.factor(title)+as.factor(student)+as.factor(year)+
as.factor(group)*size)
Anova(lm3.2,type="III")
summary(lm3.2)
```

- 实验

3.6 模型选择与预测

选择哪些自变量?

- 假设检验：选择显著水平在 0.1 或者 0.05 下能显著的影响因变量的，选择哪个显著水平合适?
- 用 AIC, BIC 进行模型选择

$$AIC = n\left\{\log(\frac{RSS}{n} + 1 + \log(2\pi))\right\} + 2 * (df + 1)$$
$$BIC = n\left\{\log(\frac{RSS}{n} + 1 + \log(2\pi))\right\} + \log(n) * (df + 1)$$

- 消耗的自由度：
  - 截距：1
  - 职称:3-1
  - 性别：2-1
  - 学生类别：3-1
  - 年份：3-1
  - 学期：2-1
  - 班级规模:2-1
  - 学生人数：1
  - 学生人数 * 班级规模: 2-1
  - 总计:12
- 手动计算全模型的 AIC,BIC，对比 R 软件计算:

```
    lm3.1=lm(score~as.factor(title)+as.factor(gender)+as.factor(student)
+as.factor(year)+as.factor(semester)+as.factor(group)*size)
Anova(lm3.1,type="III")

    AIC(lm3.1)
    AIC(lm3.1,k=log(length(score)))
```

- 跳出显著的自变量，并计算 AIC,BIC

```
    lm3.2=lm(score~as.factor(title)+as.factor(student)+as.factor(year)+
as.factor(group)*size)
Anova(lm3.2,type="III")

AIC(lm3.2)
AIC(lm3.2,k=log(length(score)))
```

AIC,BIC 均变小，说明 lm3.2 优于 lm3.1

- 对更多的模型进行比较：

```
    lm.aic=step(lm3.1,trace=F)
Anova(lm.aic,type="III")

lm.bic=step(lm3.1,k=log(length(score)),trace=F)
Anova(lm.bic,type="III")
```

结果相同

- 预测

```
    0=read.csv("/media/zheqng/Seagate Backup Plus Drive/zheqng@nwu/文档/
a0$group=1*(a0$size<=20)

score.hat=predict(lm.aic,a0)
a0$score.hat=score.hat
```

3.7 更科学的绩效评估

目标：利用合理的回归模型对原始的教评成绩予以调整。(实际上就是模型中的残差 ε)

- ```
  summary(lm.aic)
  ```

- 计算
  - 女，副教授，2002, 秋，114 人，MBA,3.175
  - 男，副教授，2002, 球，92 人，研究生，3.489

- 预测，实际就是残差：

```
a$adj.score=lm.aic$residuals
a[c(1:10),]
```
```

4 0-1 变量的回归模型

- 前三章讲述了回归分析, 方差分析和协方差分析, 它们的区别在于自变量类型不同, 但因变量都是连续的.

表 1

模型	自变量类型	因变量类型
回归分析	连续型	连续型
方差分析	离散型	连续型
协方差分析	连续 + 离散型	连续型

- 本章要处理的因变量是离散的，只有 0-1 状态，它代表事物的状态变量 (因变量)，以及影响它的因素 (自变量) 例如
  1. 银行通过财务信息预测是否破产
  2. 保险公司通过驾驶员的驾驶记录预测来年出险的可能型
  3. 信用卡经理希望通过财务和还款记录预测持卡人是否诚信.
- 因变量有如下特征:
  - 只有 0,1 两种状态
  - 变量值没有数值意义
  - 0,1 代表实物的两种状态

4.1 案例介绍

股票市场的特殊处理 (Special treatment,ST)

- ST 是特殊处理（Special Treatment）的缩写，是我国股票市场一项特有的，旨在保护投资者利益的政策。
- 如果上市公司的财务数据出现异常，则证监会将对其进行特殊处理，以便对投资者进行警示。

- 其表现特征就是在其股票名称前冠以“ST”字样
- 具体来说, 9.2.1 上市公司出现以下情形之一的，为财务状况异常：
  - 最近两个会计年度的审计结果显示的净利润均为负值;
  - 最近一个会计年度的审计结果显示其股东权益低于注册资本，即每股净资产低于股票面值;
  - 注册会计师对最近一个会计年度的财务报告出具无法表示意见或否定意见的审计报告;
  - 最近一个会计年度经审计的股东权益扣除注册会计师、有关部门不予确认的部分，低于注册资本;
  - 最近一份经审计的财务报告对上年度利润进行调整，导致连续两个会计年度亏损

研究问题与因变量

- 研究目标：从投资人的角度来看，财务报表分析能否帮助预测什么特点的公司容易被 ST, 从而避免投资损失?
- 因变量：三年以后是否被宣布 ST

自变量

- ARA：应收账款与总资产的比例，用于衡量盈利质量。
- ASSET：对数变换后的资产规模，用于反映公司规模。
- ATO：资产周转率，用于度量资产利用效率。
- GROWTH：销售收入增长率，用于反映成长潜力。
- LEV：负债资产比率，用于反映债务状况。
- ROA：资产收益率，用于度量盈利能力。
- SHARE：最大股东的持股比率，用于反映股权结构。

图 1: 共 1430 个样本，来自 1999 年的样本有 684 个，来自 2000 年的有 746 个。

4.2 基本描述

图形选择

- 对于 0-1 类型变量，散点图没有意义，盒状图是最为适用的。

- 有理由怀疑这些差异较大的变量即是影响因变量 ST 的主要因素。

4.3 单变量逻辑回归

引入

- 如果用线性回归：

$$ST = \alpha + \beta \times LEV + \varepsilon$$

- 右边是连续型，左边是离散型，矛盾！
- 将 ST 转化为一个连续型指标考虑。
- 假设存在一个为"ST 可能性" 的概念性指标，用来表示公司被 ST 的可能性
- 当"ST 可能性" 大于某一阈值时，公司会被 ST
- 推测，当两个公司的经营状况非常接近，那么，"ST 的可能性" 也应该非常接近。
- 因此可以假设，"ST 的可能性" 为一个连续型指标，取值在正负无穷之间。
- 令 $Z =$"ST 的可能性"，则模型：

$$Z = \alpha + \beta \times LEV + \varepsilon$$

- Z 是因变量：

$$\begin{aligned} P(ST = 1) &= P(Z > c) \\ &= P(\alpha + \beta \times LEV + \varepsilon > c) \\ &= P((\alpha - c) + \beta \times LEV > -\varepsilon) \\ &= F_\varepsilon(\underbrace{(\alpha - c)}_{\beta_0} + \underbrace{\beta}_{\beta_1} \times LEV) \end{aligned}$$

- F_ε 的形式：
  - $F_\varepsilon = \Phi(t) \Rightarrow$probit 模型：

$$P(ST = 1) = \Phi(\beta_0 + \beta_1 \times LEV)$$

  - $F_\varepsilon = \frac{\exp(t)}{1+\exp(t)} \Rightarrow$ 逻辑回归：

$$P(ST = 1) = \frac{\exp(\beta_0 + \beta_1 \times LEV)}{1 + \exp(\beta_0 + \beta_1 \times LEV)}$$

- 这样，对上式两边同时作用逆函数：

$$\Phi^{-1}(P(ST=1)) = \beta_0 + \beta_1 \times LEV$$
$$logit(P(ST=1)) = \beta_0 + \beta_1 \times LEV (logit(t) = log(\frac{t}{1-t}))$$

4.4 参数估计与统计推断

广义线性模型 (general linear model, 简称 glm)：
广义线性模型是对线性模型的推广。

- 对于线性模型

$$Y = \sum_i \beta_i X_i + \varepsilon$$

  Y 具有随机性，ε 是某种噪声，一般是高斯白噪声 $\varepsilon \sim \mathcal{N}(0, \sigma^2)$,$\sum_i \beta_i X_i$ 是模型的线性部分。

- 广义线性模型希望对因变量 Y 进行推广。假设 Y 服从参数为 θ 的某个分布，利用连接函数 g 将这个分布的参数 θ 与模型的线性部分连接起来。所以，广义线性模型包含三个因素
  - 随机部分：因变量 Y 服从某个概率模型 $p(Y;\theta)$(指数族，二项分布，泊松分布。。。)
  - 线性部分：$\sum_i \beta_i X_i$
  - 连接函数 (link function)：$g(\theta) = \sum_i \beta_i X_i$

参数估计

- 利用最大似然估计进行参数估计。
- 似然函数：

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \left\{ \frac{\exp(\beta_0 + \beta_1 \times LEV_i)}{1 + \exp(\beta_0 + \beta_1 \times LEV_i)} \right\}^{ST_i} \times \left\{ \frac{1}{1 + \exp(\beta_0 + \beta_1 \times LEV_i)} \right\}^{1-ST_i}$$

- 对数似然：

$$\log L(\beta_0, \beta_1) = \sum_{i=1}^{n} (ST_i) \log \left\{ \frac{\exp(\beta_0 + \beta_1 \times LEV_i)}{1 + \exp(\beta_0 + \beta_1 \times LEV_i)} \right\} + (1 - ST_i) \log \left\{ \frac{1}{1 + \exp(\beta_0 + \beta_1 \times LEV_i)} \right\}$$

- 用梯度法求解参数 $(\hat{\beta}_0, \hat{\beta}_1)$

统计推断

- t 检验：适用于 $\beta_1 \in \mathbb{R}^1$.
  $H_0 : \beta_1 = 0$
  $H_1 : \beta_1 \neq 0$
  根据中心极限定理，当样本足够多时，有下式成立：

$$\frac{\hat{\beta}_1 - \beta_1}{var(\hat{\beta}_1)} \sim \mathcal{N}(0, 1)$$

  因此，构造统计量

$$T = \frac{\hat{\beta}_1}{var(\hat{\beta}_1)}$$

- 似然比检验 (deviance, 离差)：适用于 $\beta_1 \in \mathbb{R}^s, s > 1$:

$$\lambda = -2\{\max_{\beta_0} L(\beta_0, \beta_1 = 0) - \max_{(\beta_0, \beta_1)} L(\beta_0, \beta_1)\}$$

  这里 $\lambda \sim \chi^2(d)$

5 定序回归

表 1

模型	自变量类型	因变量类型
回归分析	连续型	连续型
方差分析	离散型	连续型
协方差分析	连续 + 离散型	连续型
0-1 回归	连续	0,1 状态
定序回归	连续	定序数据

-
- 定序数据 (ordinal data): 例如，市场调查者拿着问卷和礼品要你回答问卷，多喜欢农夫山泉?
  喜欢　无所谓　不喜欢
  数据被标记为 1,2,3
- 三种情况
  - 情形 1(有数值意义)：1 岁，2 岁，3 岁
  - 情形 2(无数值意义，无顺序意义)：红色，蓝色，绿色
  - 情形 3(无数值意义，有顺序意义): 不喜欢，无所谓，喜欢

  情形 3 是本章要研究的类型。

5.1 案例介绍

研究目标

消费者对于不同类型手机的偏好

数据来源

- 对 MBA 学生的调查

- 共 1451 个观测值

假设

- 在传统的对手机功能偏好的研究中，往往假设增加某一新功能所带来的影响同手机的其他现有功能无关 (这一假设显然不合理)
- 同样的功能在低端手机和高端手机上的结果不一样，甚至有可能有巨大差异。

变量介绍

变量类型	变量含义	变量名	变量水平
因变量	对该产品的偏好程度	score	1=根本不喜欢；2=比较不喜欢；3=一般喜欢；4=比较喜欢；5=非常喜欢
自变量	手机品牌	W1	共四种（诺基亚、摩托罗拉、三星和波导）
	有无数码相机	W2	共二种（有、无）
	能否收看电视	W3	共二种（能、不能）
	有无手写笔	W4	共二种（有、无）
	电话本能否多条记录	W5	共二种（能、不能）
	有无 MP3	W6	共二种（有、无）
	游戏数目	W7	连续型

图 1: 变量介绍

实验设计

品牌	数码相机	能否收看电视	手写笔	电话本能否多条记录	MP3	游戏数目
诺基亚	无	不能	无	能	有	3
	有	不能	有	不能	有	5
	无	能	有	不能	无	7
波导	有	能	无	能	无	3
	无	不能	无	不能	有	5
	有	不能	有	能	有	7
摩托罗拉	无	能	有	能	无	3
	有	能	无	不能	无	5
	无	不能	无	能	无	7
三星	有	不能	有	不能	无	3
	无	能	有	不能	有	5
	有	能	无	能	有	7

图 2: 实验设计

5.2 描述性分析

- 数据读入
- 列链表
- 画图

5.3 定序回归模型

- 观测数据 $(W_i, score_i)$, 其中 $score_i \in \{1, \cdots, K\}$,$W_i = (W_1, \cdots, W_7)$
- 本节先考虑单变量模型，自变量选择 $W7$ 作为研究对象。
- 线性部分是 $W7\beta_7 + \varepsilon \in \mathbb{R}$, 而因变量 $score \in \mathbb{N}$, 所以希望借助隐变量 $Z \in \mathbb{R}$ 建立自变量与因变量的关系。
- 隐变量模型：

$$Z = \beta_0 + W7 \times \beta_7 + \varepsilon \tag{5.1}$$

并且有

$$score = \begin{cases} 1 & Z < c_1 \\ 2 & c_1 \le Z < c_2 \\ 3 & c_2 \le Z < c_3 \\ 4 & c_3 \le Z < c_4 \\ 5 & c_4 \le Z \end{cases}$$

因而

$$P(score \le k) = P(Z \le c_k) \tag{5.2}$$

将5.1代入5.2有

$$\begin{aligned} P(score \le k) &= P(\beta_0 + W7 \times \beta_7 + \varepsilon \le c_k) \\ &= P((c_k - \beta_0) - W7 \times \beta_7 \ge \varepsilon) \\ &= F_\varepsilon(\underbrace{(c_k - \beta_0)}_{\alpha_k} - W7 \times \beta_7) \end{aligned}$$

这里不关心截距 α_k，只关心斜率 β_7。如果 $\beta_7 = 0$，说明自变量 $W7$ 对于因变量 $score$ 没有影响; 如果 β_7 在 α 的显著水平下没有通过显著性检验，说明自变量 $W7$ 对于因变量 $score$ 没有显著的影响，可以被忽略；如果 β_7 在 α 的显著水平下通过显著性检验，说明 $W7$ 对于 $score$ 有着显著的影响。

- F_ε 有两种选择
  - $F_\varepsilon = \Phi(t) \Rightarrow$probit 模型：

$$P(score \le k) = \Phi(\alpha_k - \beta_7 \times W7)$$

  - $F_\varepsilon = \frac{\exp(t)}{1+\exp(t)} \Rightarrow$ 逻辑回归：

$$P(score \le k) = \frac{\exp(\alpha_k - \beta_7 \times W7)}{1 + \exp(\alpha_k - \beta_7 \times W7)}$$

- 这样，对上式两边同时作用逆函数：

$$\Phi^{-1}(P(score \leq k)) = \alpha_k - \beta_7 \times W7$$
$$logit(P(score \leq k)) = \alpha_k - \beta_7 \times W7 (logit(t) = log(\frac{t}{1-t}))$$

5.4 参数估计与统计推断

参数估计

- 利用最大似然估计进行参数估计。
- 似然函数：

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \left\{ \frac{\exp(\beta_0 + \beta_1 \times LEV_i)}{1 + \exp(\beta_0 + \beta_1 \times LEV_i)} \right\}^{ST_i} \times \left\{ \frac{1}{1 + \exp(\beta_0 + \beta_1 \times LEV_i)} \right\}^{1-ST_i}$$

- 对数似然：

$$\log L(\beta_0, \beta_1) = \sum_{i=1}^{n} (ST_i) \log \left\{ \frac{\exp(\beta_0 + \beta_1 \times LEV_i)}{1 + \exp(\beta_0 + \beta_1 \times LEV_i)} \right\} + (1 - ST_i) \log \left\{ \frac{1}{1 + \exp(\beta_0 + \beta_1 \times LEV_i)} \right\}$$

- 用梯度法求解参数 $(\hat{\beta}_0, \hat{\beta}_1)$

统计推断

- t 检验：适用于 $\beta_1 \in \mathbb{R}^1$.
  $H_0 : \beta_1 = 0$
  $H_1 : \beta_1 \neq 0$
  根据中心极限定理，当样本足够多时，有下式成立：

$$\frac{\hat{\beta}_1 - \beta_1}{var(\hat{\beta}_1)} \sim \mathcal{N}(0, 1)$$

  因此，构造统计量

$$T = \frac{\hat{\beta}_1}{var(\hat{\beta}_1)}$$

- 似然比检验 (deviance, 离差)：适用于 $\beta_1 \in \mathbb{R}^s, s > 1$:

$$\lambda = -2\{\max_{\beta_0} L(\beta_0, \beta_1 = 0) - \max_{(\beta_0, \beta_1)} L(\beta_0, \beta_1)\}$$

  这里 $\lambda \sim \chi^2(d)$

5.5 多变量逻辑回归

```
library(MASS)
probit0=polr(as.factor(score)~1,method="probit",Hess=T)
probit1=polr(as.factor(score)~W1+W2+W3+W4+W5+W6+W7,method="probit",Hess=T)
anova(probit0,probit1)

logit0=polr(as.factor(score)~1,method="logistic",Hess=T)
logit1=polr(as.factor(score)~W1+W2+W3+W4+W5+W6+W7,method="logistic",Hess=T)
anova(logit0,logit1)

library(car)
Anova(probit1,type="III")
summary(probit1)

Anova(logit1,type="III")
summary(logit1)

probit2=polr(as.factor(score)~W1+W2+W3+W4+W5+W6,method="probit",Hess=T)
summary(probit2)

logit2=polr(as.factor(score)~W1+W2+W3+W4+W5+W6,method="logistic",Hess=T)
summary(logit2)
```

5.6 模型选择

模型选择采用 AIC 和 BIC

$$AIC = deviance + 2 \times df$$

$$BIC = deviance + \log(n) \times df$$

```
logit.aic=step(logit1,trace=F)
summary(logit.aic)

logit.bic=step(logit1,trace=F,k=log(length(a[,1])))
summary(logit.bic)
```

```
probit.aic=step(probit1,trace=F)
summary(probit.aic)

probit.bic=step(probit1,trace=F,k=log(length(a[,1])))
summary(probit.bic)
```

5.7 预测与评估

```
summary(logit.aic)
a0$score.hat=predict(probit.aic,a0)
```

6 泊松回归

- 定序数据 (无数值意义，有顺序意义)：1= 不喜欢，2= 无所谓,3= 喜欢
- 泊松回归 (有数值意义): 例如，某顾客某月内光顾超市的次数，1,2,3 次

6.1 案例介绍

- 数据来源
- 共 3995 个样本，因变量：freq0: 基准月份频率
- 自变量如下图

解释性变量

变量名称	作用	实际意义
freq0	因变量	第 0 月光顾超市的频数
freq1	自变量	第-1 月光顾超市的频数
freq2	自变量	第-2 月光顾超市的频数
freq3	自变量	第-3 月光顾超市的频数
exp1	自变量	第-1 月的消费金额
exp2	自变量	第-2 月的消费金额
exp3	自变量	第-3 月的消费金额

图 1

6.2 描述性分析

盒状图

```
boxplot(freq1~freq0,xlab="freq0",ylab="freq1")
boxplot(exp1~freq0,xlab="freq0",ylab="exp1")

par(mfrow=c(2,3))
boxplot(freq1~freq0,xlab="freq0",ylab="freq1",main="第-1月")
boxplot(freq2~freq0,xlab="freq0",ylab="freq2",main="第-2月")
boxplot(freq3~freq0,xlab="freq0",ylab="freq3",main="第-3月")
boxplot(exp1~freq0,xlab="freq0",ylab="exp1",main="第-1月")
boxplot(exp2~freq0,xlab="freq0",ylab="exp2",main="第-2月")
boxplot(exp3~freq0,xlab="freq0",ylab="exp3",main="第-3月")
par(mfrow=c(1,1))
```

6.3 泊松回归

泛化线性模型 glm:

- 线性部分：$\beta_0+\beta_1\times freq1+\beta_2\times freq2+\beta_3\times freq3+\beta_4\times exp1+\beta_5\times exp2+\beta_6\times exp3$
- 随机项：$freq0\sim\mathcal{P}(\lambda)$
- 连接函数:$g^{-1}(\lambda)=\beta_0+\beta_1\times freq1+\beta_2\times freq2+\beta_3\times freq3+\beta_4\times exp1+\beta_5\times exp2+\beta_6\times exp3$

两点说明：

- 由于 $\lambda>0$ 只能取正值，所以选择连接函数为 $g^{-1}(\lambda)=\log(\lambda)$
- 对因变量建立泊松模型，有着三个假设：
  - 平稳性
  - 增量独立性
  - 无记忆性

  而前两个假设恰好就是泊松分布的两大假设，在适当的条件下，可以证明满足这一性质的随机变量，分布只能是泊松分布 (参见相关概率论的知识)。

这样，我们有

- 因变量的分布是：
$$P(freq0=k)=\frac{\lambda^k}{k!}\exp(-\lambda)$$
其中 $\lambda=E[freq0]$ 是顾客平均每月光顾超市的次数。

- 通过连接函数建立的模型为：

$$\log\lambda(x)=\beta_0+\beta_1\times freq1+\beta_2\times freq2+\beta_3\times freq3+\beta_4\times exp1+\beta_5\times exp2+\beta_6\times exp3+\varepsilon$$

6.4 参数估计与统计推断

参数估计 (最大似然估计)

似然函数为

$$L(\beta_0,\boldsymbol{\beta})=\prod_{i=1}^{n}\frac{\lambda(x_i)^{k_i}}{k_i!}\exp\left(-\lambda(x_i)\right)$$

$$\log\lambda(\boldsymbol{x}_i)=\beta_0+\beta_1\times x_{i1}+\cdots+\beta_p\times x_{ip}$$

其中，$\boldsymbol{x}_i=(x_{i1},\cdots,x_{ip}),\ i=1,\cdots,n$ 对应着 (freq1,freq2,freq3,exp1,exp2,exp3)

最大似然相当于对对数似然求最大化：

$$\log L(\beta_0,\boldsymbol{\beta})=\sum_{i=1}^{n}k_i\log\lambda(\boldsymbol{x}_i)-\log k_i!-\lambda(\boldsymbol{x}_i)$$

统计推断

- 单参数统计推断:(t-检验)：

  由中心极限定理：

$$\frac{\hat{\beta}_j-\beta_j}{\sqrt{var(\hat{\beta}_j)}}\sim\mathcal{N}(0,1),\ (j=0,1,\cdots,p)$$

  构造 t 统计量为

$$T_j=\frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)}$$

- 多参数推断:(似然比检验)

$$\lambda=-2\times(\max_{\beta_0}L(\beta_0,\boldsymbol{\beta}=0)-\max_{\beta_0,\boldsymbol{\beta}}\log L(\beta_0,\boldsymbol{\beta}))$$

$$\lambda\sim\chi^2(p)$$

```
pos0=glm(freq0~1,family=poisson())
pos1=glm(freq0~freq1+freq2+freq3+exp1+exp2+exp3,family=poisson())
anova(pos0,pos1)
1-pchisq(2146.0,df=6)

library(car)
Anova(pos1,type="III")

summary(pos1)
```

6.5 模型选择与预测

模型选择

$$AIC = deviance + 2 \times df$$

$$BIC = deviance + \log(n) \times df$$

```
pos.aic=step(pos1,trace=F)
summary(pos.aic)

pos.bic=step(pos1,trace=F,k=log(length(a[,1])))
summary(pos.bic)
```

模型预测

$$\lambda(\boldsymbol{x}_i) = \exp(\beta_0 + \beta_1 \times x_{i1} + \cdots + \beta_p \times x_{ip})$$

误差：

$$\sqrt{\frac{1}{m}\sum_{i=1}^{m}(k_i^{true} - k_i^{pred})^2}$$

7 生存分析模型

- 数据类型，特别是因变量类型，决定着统计分析方法
- 连续型因变量 $\rightarrow$ 回归方法 (线性回归，方差分析，协方差分析)
- 离散型因变量 $\rightarrow$ 广义线性回归 (逻辑回归，泊松回归，定序回归)
- 生存数据类型：例如
  1. 癌症病人：从确诊到死亡的时间
  2. 公司：从成立到破产的时间
  3. 品牌：从上市到消失的时间
- 在有限的观测时间内观测不到病人死亡 (数据特点)，因而不能单纯根据因变量 Time 是连续型，就选用回归分析，需要建立新的模型

7.1 案例介绍

- 数据共 65 个病人,1975 年收集的，48 人在研究期间死去，17 人活过了研究期限。
- 变量如下

7.2 生存函数

- 引入：例如，对于生存数据，在计算均值时，被截断的数据不知道精确值。把这部分数据直接扔掉，利用剩下的数据计算均值造成了数据的浪费; 另一方面，利用这些数据计算均值又造成了均值估计偏小。实际上，对于某个截断数据，我们确切的知道他至少生存了多长时间，这个数据是可以被利用的。
- 生存函数：

$$S(t) = P(T > t)$$

变量介绍

变量名称	解释意义
Time	从确诊到死亡的生存时间（单位：月）
VStatus	生存状态（0=生存；1=死亡）
HGB	确诊时血色素（Hemoglobin）含量
Platelet	确诊时血小板（0=不正常；1=正常）
Age	确诊时年龄（单位：年）
LogWBC	对数变换后白细胞（White Blood Ceil）含量
LogPBM	对数变换后骨髓中血浆细胞含量
Protein	确诊时血蛋白含量
SCalc	确诊时血清钙含量

图 1

这是随机变量 T 的分布，表示该样本至少存活时间为 t 的概率。
对于被截断数据，估计生存函数为

$$\tilde{S}(t) = \frac{1}{n}\sum_{i=1}^{n} I(t_i > t)$$

但这里存在着被截断的数据，因此需要另寻方法。

- 符号：记第 i 个个体 $\begin{cases} t_i & \text{未截断} \\ t_i+ & \text{截断 (说明生存时间} > t_i) \end{cases}$

```
library(survival)
a=a[order(a$Time),]
Surv(a$Time,a$VStatus)
```

- 在险者个数 (number of case at risk) $r(t)$: 样本确知存活时间为 t 的个数
- 事件个数 (number of events) $d(t)$: 时间点 t 的死亡个数
- Kaplan-Meier 估计

$$P(T > t|T \geq t) \approx 1 - \frac{d(t)}{r(t)}$$

由于 $d(t)$ 在大量的时间点上均为 0，只考虑 $E = \{t : d(t) \neq 0\}$ 这样估计生存函数为：

$$S(t) = P(T > t) = \prod_{t_i \leq t, t_i \in E} P(T > t_i | T \geq t_i) \approx \prod_{t_i \leq t, t_i \in E} \left(1 - \frac{d(t_i)}{r(t_i)}\right)$$

方差

$$var(\tilde{S}(t)) \approx \tilde{S}^2(t) \sum \frac{d(t_i)}{r(t_i)(r(t_i) - d(t_i))}$$

```
summary(survfit(Surv(a$Time,a$VStatus)~1))
plot(survfit(Surv(a$Time,a$VStatus)~1))
```

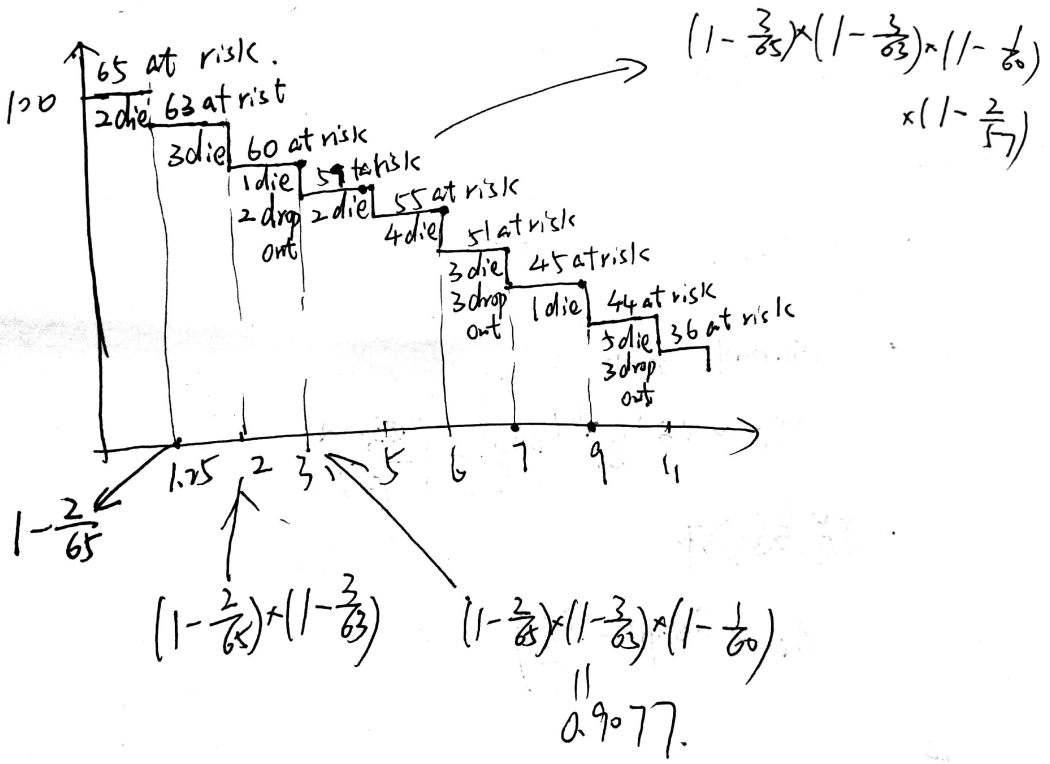


图 2

7.3 描述性分析

用 survfit 函数

```
status=1*(a$HGB>median(a$HGB))
plot(survfit(Surv(a$Time,a$VStatus)~status),col=c(1,2), lty=c(1,2))
legend(40,1,c("HGB<Median","HGB>Median"),col=c(1,2),lty=c(1,2))

plot(survfit(Surv(a$Time,a$VStatus)~a$Platelet),col=c(1,2),lty=c(1,2))
legend(40,1,c("Abnormal","normal"),col=c(1,2),lty=c(1,2))

age.group=1*(a$Age>median(a$Age))
plot(survfit(Surv(a$Time,a$VStatus)~age.group),col=c(1,2),lty=c(1,2))
legend(40,1,c("Age<60","Age>60"),col=c(1,2),lty=c(1,2))

par(mfrow=c(2,2))
status=1*(a$LogWBC>median(a$LogWBC))
plot(survfit(Surv(a$Time,a$VStatus)~status),col=c(1,2),lty=c(1,2))
legend(30,1,c("LogWBC<Median","LogWBC>Median"),col=c(1,2),lty=c(1,2))

status=1*(a$LogPBM>median(a$LogPBM))
plot(survfit(Surv(a$Time,a$VStatus)~status),col=c(1,2),lty=c(1,2))
legend(30,1,c("LogPBM<Median","LogPBM>Median"),col=c(1,2),lty=c(1,2))

status=1*(a$Protein>median(a$Protein))
plot(survfit(Surv(a$Time,a$VStatus)~status),col=c(1,2),lty=c(1,2))
legend(30,1,c("Protein<Median","Protein>Median"),col=c(1,2),lty=c(1,2))

status=1*(a$SCalc>median(a$SCalc))
plot(survfit(Surv(a$Time,a$VStatus)~status),col=c(1,2),lty=c(1,2))
legend(30,1,c("SCalc<Median","SCalc>Median"),col=c(1,2),lty=c(1,2))
par(mfrow=c(1,1))
```

7.4 加速死亡模型

- 理想的情况：t_i 均被观测到，对第 i 个病人，知道和 t_i 相关的解释性变量 $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})^T$, 建立模型：

$$\log t_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i$$

- ε_i 是服从某种分布的残差项，也可以认为是在没有协变量影响 ($\boldsymbol{x}_i = 0$) 下的对数生存时间。假设：
  - 正态分布：$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$
  - Weibull 分布：$\varepsilon_i \sim \alpha\beta^\alpha \varepsilon_i^{\alpha-1} \exp\left(-\left(\frac{\varepsilon_i}{\beta}\right)^\alpha\right)$
  - 指数分布：$\varepsilon_i \sim \beta \exp\left(-\frac{\varepsilon_i}{\beta}\right)$

7.5 Cox 风险模型

- 风险 (hazard) 定义为：

$$h(t) = \lim_{\delta t \to 0} \frac{P(t \le t_i \le t + \Delta t | t_i \ge t)}{\Delta t} = -\frac{f(t)}{S(t)}$$

  $f(t)$ 是生存函数 $S(t)$ 的导数，即生存时间 t_i 的密度函数
  $h(t)$ 度量了某个个体已经存活了时间 t，立刻死亡的可能性。理论上

$$S(t) = \exp(-\int_0^t h(s)ds)$$

- Cox 的等比例风险模型 (Cox's proportional Hazard Model)

$$h(t) = h_0(t) \exp(\boldsymbol{\beta}^T \boldsymbol{x})$$

  - $h_0(t)$ 称为基准风险函数 (Baseline Hazard Function)，表示在没有协变量的影响 (即 x=0) 下，一个病人的风险函数形式
  - 给定协变量 $\boldsymbol{x}$ 的影响，基准风险函数 $h_0(t)$ 被等比例扩大了 $\exp(\boldsymbol{\beta}^T \boldsymbol{x})$ 倍。
  - 恰当的选择 $h_0(t)$ 的形式，Cox 模型包含了 Weibull 加速死亡模型。
  - Cox 模型更加灵活，可以不对 $h_0(t)$ 作任何假设。

- 偏最大似然估计
  - 定义 $R(t)$ 是集合：包含所有存活时间至少为 t 的个体
  - 在存活时间 $\ge t_i$ 的个体中，个体死亡的可能性

$$\frac{h_0(t_i) \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}{\sum_{j \in R(t_i)} h_0(t_i) \exp(\boldsymbol{\beta}^T \boldsymbol{x}_j)} = \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}{\sum_{j \in R(t_i)} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_j)}$$

– 偏似然估计

$$\prod_{t_i \in E} \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}{\sum_{j \in R(t_i)} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_j)}$$

8 自回归模型

- 横截面数据：多个个体，单个观测 (例如，第一张的盈利预测，同时考虑成百上千的公司，每个公司是一个个体，每个公司某年的财务信息，是单个观测。)
- 时间序列数据：一个个体随时间的推移有多个观测 (例如，一个国家的 GDP，每年的 GDP 取值，形成不同观测)
- 对于横截面数据，不同数据来自不同个体，因此可以假设不同数据独立。
- 对于时间序列数据，不同数据来自同一个体，因此具有很强的相关性。

8.1 案例介绍

- 研究某个国家的失业率
- 失业率是衡量一个国家或地区就业状况的最重要的指标之一，也是反映社会稳定性的重要指标之一。对于相关机构，特别是政府职能部门，意义重大。
- 1948 年 1 月至 2006 年 12 月，该国家各个月份的失业率 (%)
- 数据受月份影响很大，本数据是通过季节性调整后的失业率。

关心的问题：

- 失业率的变化规律。例如当月的失业率同过去几个月的失业率是否有关？什么关系？
- 基于失业率的变化规律作预测。

8.2 时间序列的平稳性

- 变化巨大的数据，如何建立良好的预测模型？
- 为什么统计模型具有预测能力？
- 本案例，用 1948 年至 2006 年 58 年的数据预测 2007 年的失业率。必须假设 2007 年的失业率的变化规律同 1948 年-2006 年的变化规律具有某种相似性

- 用平稳性来定义这种相似性。

定义 (严平稳)
对于一个任意的整数 $k > 0$, 随机变量 $(X_{t_0+1}, X_{t_0+2}, \cdots, X_{t_0+k})$ 的分布同时间点 t_0 无关。

换句话说，对任意的 $t_1 \neq t_2$,$(X_{t_1+1}, \cdots, X_{t_1+k})$ 与 $(X_{t_2+1}, \cdots, X_{t_2+k})$ 的分布相同。
引入平稳性的概念有两方面的意义：

- 容易量化历史数据 $\{x_1, \cdots, x_T\}$ 中相邻观测 (x_t, x_{t+1}), $(t = 1, \cdots, T-1)$ 的回归关系 (例如，$E(x_{t+1}|x_t) = 0.5x_t$);
- 可以作预测 (例如，$x_{T+1} = 0.5x_T$)

举例

- 非平稳性的例子: 本案例，失业率具有明显的下降趋势，因而非平稳; 随机游走，图 8-2;
- 平稳性的例子: 图 8-3, 没有明显的变化趋势; 把失业率数据作对数差的变换，$r_t = \log x_t - \log x_{t-1}$, 序列 $\{r_t\}$ 是平稳序列。

8.3 基本描述

盒状图

8.4 自相关系数

- 时间序列的一个重要特征就是未来数据同历史数据的相关性，称自相关性 (autocorrelation)
- 对横截面数据，通过相关系数 (correlation coefficient) 来描述两个变量之间的线性相关性
- 对时间序列，$\rho_k = corr(X_t, X_{t+k})$, $k > 0$ 度量相距为 k 个单位时间的两个时间序列数据的相关性，称 ρ_k 为 k 阶自相关系数 (ACF):
  - $|\rho_k|$ 大，表示当月失业率与上月失业率相关性很强。
  - $|\rho_k|$ 小，表示当月失业率与上月失业率相关性很弱
  - $\rho_k > 0$，表示上月失业率上升，会带来当月失业率上升
  - $\rho_k < 0$, 表示上月失业率上升，会带来当月失业率下降。

```
acf(a$rate)
acf(r)
```

8.5 自回归模型及其平稳性

- 模型:

$$r_t = a_0 + a_1 r_{t-1} + \cdots + a_p r_{t-p} + \varepsilon_t$$

  $\boldsymbol{a} = (a_0, a_1, \cdots, a_p)$ 是自回归系数,p 是模型阶数。ε_t 是方差为 σ^2 的白噪声。

- 如果一个时间序列模型是成功的，那么它所分离出的残差就不应该掺杂任何时间序列特征，所有的时间序列特征都应该被模型充分利用，以提高预测精度。

- 注意，$x_t, x_{t'}$ 相关;$\varepsilon_t, \varepsilon_{t'}$ 独立。

在参数估计之前，考虑什么样的自回归模型是平稳的?

- $r_t = r_{t-1} + \varepsilon_t,\ \ var(\varepsilon_t) = 1$, 非平稳
- $r_t = \sqrt{2} r_{t-1} + \varepsilon_t,\ \ var(\varepsilon_t) = 1$，非平稳
- $r_t = 2^{-1/2} r_{t-1} + \varepsilon_t,\ \ var(\varepsilon_t) = 1,\ \ E(\varepsilon_t) = 0$，平稳

理论上，对 $r_t = a_0 + a_1 r_{t-1} + \varepsilon_t$, 平稳的充要条件是 $|a_1| < 1$;
对于一般的 $r_t = a_0 + a_1 r_{t-1} + \cdots + a_p r_{t-p}$, 平稳的充要条件是 $1 - a_1 z - \cdots z_p z^p = 0$ 的复数根全部在单位圆外。

8.6 模型估计与选择

参数估计 (最小二乘法)

$$\hat{\phi} = \arg\min_{\phi} \sum_{t=p+1}^{T} (r_t - a_0 - \cdots - a_p r_{t-p})^2$$

```
ar(r, aic=F, order=4)
```

模型选择 (AIC,BIC)

```
fit=ar(r)
plot(0:28, fit$aic, type="b")
```

8.7 模型诊断

检查模型分离出的白噪声是否具有时序特征：

- 检查 ε_t 的相关系数
- 异常值检验
- 正态性检验

8.8 模型预测

- 单点预测: 利用第 1 600 个数据建立 $AR(2)$ 模型，对第 601 个数据作预测。
- 相对预测误差: 利用第 1 600 个数据建立 $AR(2)$ 模型，对第 601 707 个数据作预测:

$$RPE = \frac{\sum_{t=601}^{707}(r_t^{true} - r_t^{pred})^2/107}{\sum_{t=601}^{707}(r_t^{true} - \overline{r}_t^{true})^2/107}, \quad \overline{r}_t^{true} = \frac{\sum_{t=601}^{707} r_t^{true}}{107}$$

  与均值预测相比，AR(2) 有多大改进。