第 2 章　一元线性回归

李　杰

数据科学学院，浙江财经大学

2017 年 10 月 22 日

内容提要

2.1 一元线性回归模型

2.2 参数 β_0, β_1 的估计

2.3 最小二乘估计的性质

2.4 回归方程的显著性检验

2.5 残差分析

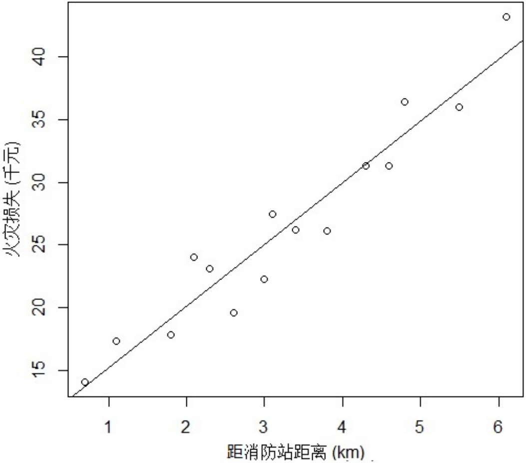
2.6 回归系数的区间估计

2.7 预测与控制

2.8 小结与评注

☞ 一元线性回归模型的实际背景

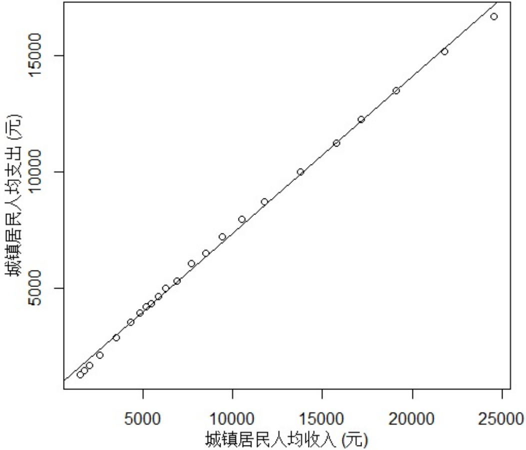
- 在实际问题的研究中，常需要研究某一现象与影响它的某一最主要因素的关系.
- 例: **火灾损失** (*Loss*) v.s. 火灾发生地与最近消防站**距离** (*Distance*)



- 代码示例

```
fire <- data.frame(distance=numeric(0), loss = numeric(0))
                                        # 生成一个数据框
fire <- edit(fire)                      # 编辑数据
plot(loss~distance,data=fire)           # 绘制"损失~距离"的散点图
abline(lm(loss~distance,data=fire)) # 添加回归线
write.csv(fire,"d:/regress/fire.csv",row.names=F)
                                        # 保存数据文件
```

- 城镇人均支出 (*expend*) *v.s.* 人均收入 (*income*)



- 代码示例

```
expend <- data.frame(year=numeric(0),expend=numeric(0),
                                               income=numeric(0))
expend <- edit(expend)
plot(expend~income, data=expend)
abline(lm(expend~income, data=expend))
write.csv(expend,"d:/regress/expend.csv",row.names=F)
```

☞ 一元线性回归模型的数学形式

- **一元线性理论回归模型**

$$y = \underline{\beta_0 + \beta_1 x} + \underline{\varepsilon} \tag{1}$$

  - y: 被解释变量 (因变量);
  - x: 解释变量 (自变量);
  - β_0: 回归截距;
  - β_1: 回归系数;
  - ε: 随机误差 (扰动项, 误差项), 常假定

$$E(\varepsilon) = 0, \quad \text{Var}(\varepsilon) = \sigma^2 \tag{2}$$

  - **理论回归模型 (1) 的含义:** 被解释变量 y 的变化 (波动) 因两部分引起, 一部分由解释变量 x 引起的线性变化 $\beta_0 + \beta_1 x$, 另一部分由随机因素 ε 引起的.

- **总体回归方程**

  - 在假设 (2) 下, 有

$$E(y|x) = \beta_0 + \beta_1 x \tag{3}$$

式 (3) 称为总体回归方程, 也简记为 $E(y)$.

- **样本回归模型**
  - 设 (x_1, y_1), (x_2, y_2), $\cdots$, (x_n, y_n) 是来自总体 (x, y) 的样本, 则样本满足

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad (i = 1, 2, \cdots, n) \tag{4}$$

称式 (4) 为样本回归模型, 由式 (2) 可知

$$\mathrm{E}(\varepsilon_i) = 0, \quad \mathrm{Var}(\varepsilon_i) = \sigma^2 \quad (i = 1, 2, \cdots, n) \tag{5}$$

以及

$$\mathrm{E}(y_i|x_i) = \beta_0 + \beta_1 x_i, \quad \mathrm{Var}(y_i|x_i) = \sigma^2 \quad (i = 1, 2, \cdots, n) \tag{6}$$

  - 假设: 通常假定 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n \sim i.i.d.$, 故 $y_1, y_2, \cdots, y_n$ 相互独立但不同分布 (why?).

- **一元线性回归分析的任务**
  - 基于样本观测值 (x_1, y_1), (x_2, y_2), $\cdots$, (x_n, y_n) 估计β_0, β_1, 得到一元线性经验回归方程

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{7}$$

- 在实际问题的研究中，为了方便地对参数做区间估计和假设检验，常假定

$$\varepsilon \sim N(0, \sigma^2) \tag{8}$$

故有

$$\varepsilon_i \sim N(0, \sigma^2) \tag{9}$$
$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \quad (i = 1, 2, \cdots, n) \tag{10}$$

- **一元线性回归模型的矩阵表示**
  - ✧ 令

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \tag{11}$$

则有

$$\begin{cases} Y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0} \\ \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n \end{cases} \tag{12}$$

☞ **普通最小二乘估计** (ordinary least square estimation, OLSE)

- **基本思想:** 对每个样本观测值 (x_i, y_i), 考虑观测值 y_i 与其回归值 $E(y_i|x_i) = \beta_0 + \beta_1 x_i$ 的离差, 综合考虑 n 个离差值, 定义离差平方和

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - E(y_i|x_i)]^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{13}$$

所谓最小二乘法, 就是寻找 β_0, β_1 的估计值 $\hat{\beta}_0, \hat{\beta}_1$ 使得式 (13) 定义的离差平方和最小, 故满足

$$Q(\hat{\beta}_0, \hat{\beta}_1) = \min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{14}$$

的 $\hat{\beta}_0, \hat{\beta}_1$ 被称为 β_0, β_1 的最小二乘估计 (OLSE).

- **回归拟合值**. 称

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

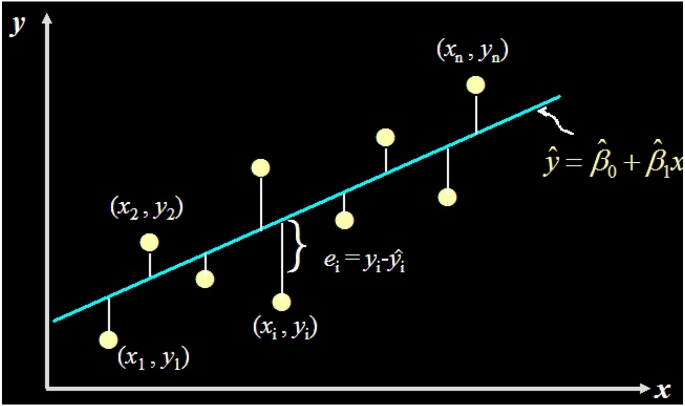
为 $E(y_i|x_i)$ 的回归拟合值 (回归值、拟合值).

- **残差**. 称

$$e_i = y_i - \hat{y}_i$$

为 y_i 的残差. 残差 e_i 常作为误差项 ε_i 的估计值.

- 经验回归方程



- **最小二乘估计的推导**

$$\min_{\beta_0,\beta_1} Q(\beta_0,\beta_1) = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

【解】由极值原理，可得

$$\begin{cases} \left.\dfrac{\partial Q}{\partial \beta_0}\right|_{\substack{\beta_0=\hat\beta_0 \\ \beta_1=\hat\beta_1}} = -2\sum_{i=1}^{n}\left(y_i - \hat\beta_0 - \hat\beta_1 x_i\right) = 0 \\ \left.\dfrac{\partial Q}{\partial \beta_1}\right|_{\substack{\beta_0=\hat\beta_0 \\ \beta_1=\hat\beta_1}} = -2\sum_{i=1}^{n}\left(y_i - \hat\beta_0 - \hat\beta_1 x_i\right) x_i = 0 \end{cases}$$

整理后得**正规方程组** (regular equations)

$$\begin{cases} n\hat\beta_0 + \left(\sum_{i=1}^{n} x_i\right)\hat\beta_1 = \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i\right)\hat\beta_0 + \left(\sum_{i=1}^{n} x_i^2\right)\hat\beta_1 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

求解正规方程组，可得

$$\begin{cases} \hat{\beta}_1 = \dfrac{\sum\limits_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2} \\ \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \end{cases}$$

其中

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- **注:**

  ① 记

$$L_{xx} = \sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2$$

$$L_{xy} = \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

  则有

$$\begin{cases} \hat{\beta}_1 = \dfrac{L_{xy}}{L_{xx}} \\ \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \end{cases}$$

② $\hat{\beta}_1$ 还可等价地表示为

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

或

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

③ 由 $\hat{\beta}_0$ 可知

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}$$

即经验回归方程一定经过样本均值点 $(\bar{x}, \bar{y})$.

- **残差的性质**

$$\sum_{i=1}^{n} e_i = 0, \quad \sum_{i=1}^{n} e_i x_i = 0$$

☞ 最大似然估计 (maximum likelihood estimation, MLE)

- 由误差项的正态性假设 $\varepsilon_i \sim N(0, \sigma^2)$ 可知

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

故 y_i 的密度函数为

$$f_i(y_i) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}[y_i - (\beta_0 + \beta_1 x_i)]^2\right\} \quad (i = 1, 2, \cdots, n)$$

所以 $(y_1, y_2, \cdots, y_n)$ 的似然函数为

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} f_i(y_i) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i - \beta_0 + \beta_1 x_i]^2\right\}$$

相应的对数似然函数为

$$\ln(L) = -\frac{n}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i - \beta_0 + \beta_1 x_i]^2$$

导出的最大似然估计量为

$$\begin{cases} \hat{\beta}_1 = \frac{L_{xy}}{L_{xx}} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{y}_i]^2 = \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i]^2 \end{cases}$$

- **注:**
  - 使用最大似然估计法的前提是已知误差项的分布类型.
  - $y_1, y_2, \cdots, y_n$ 虽然不同分布, 但在独立性假设下仍可方便地导出似然函数.
  - σ^2 的无偏估计为

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}[y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i]^2$$

☞ 代码示例

```
fire_lm <- lm(loss~distance,data=fire) # 最小二乘回归估计
summary(fire_lm)          # 查看估计结果

coefficients(fire_lm)                  # 取回归系数估计值
confint(fire_lm)                       # 取回归系数区间估计，缺省水平 95%
confint(fire_lm,parm=2)                # 取第2个系数的区间估计，缺省水平95%
confint(fire_lm,parm=2,level=0.99)     # 取第2个系数的区间估计，缺省水平99%

fitted(fire_lm)                        # 取估计的拟合值
residual(fire_lm)                      # 取估计的残差

vcov(fire_lm)                          # 取参数估计量的协方差矩阵
```

☞ **最小二乘估计量的线性性**

估计量的线性性

最小二乘估计量 $\hat{\beta}_0, \hat{\beta}_1$ 是关于随机变量 $y_1, y_2, \cdots, y_n$ 的线性函数.

【证】令 $d_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$，则

$$\hat{\beta}_1 = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})y_i}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} = \sum_{i=1}^{n} \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} y_i = \sum_{i=1}^{n} d_i y_i$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \sum_{i=1}^{n} \left(\frac{1}{n} - d_i \bar{x} \right) y_i$$

☞ 最小二乘估计量的无偏性

估计量的无偏性

$\mathrm{E}(\hat{\beta}_0)=\beta_0, \quad \mathrm{E}(\hat{\beta}_1)=\beta_1.$

【证】

$$\mathrm{E}(\hat{\beta}_1)=\sum_{i=1}^{n}\frac{x_i-\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\mathrm{E}(y_i)=\sum_{i=1}^{n}\frac{x_i-\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}(\beta_0+\beta_1 x_i)=\beta_1$$

$$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}=(\beta_0+\beta_1\bar{x}+\bar{\varepsilon})-\hat{\beta}_1\bar{x}=\beta_0+(\beta_1-\hat{\beta}_1)\bar{x}+\bar{\varepsilon}$$

$$\mathrm{E}(\hat{\beta}_0)=\beta_0+\mathrm{E}[(\beta_1-\hat{\beta}_1)\bar{x}]+\mathrm{E}(\bar{\varepsilon})=\beta_0+\mathrm{E}[(\beta_1-\hat{\beta}_1)]\bar{x}=\beta_0$$

☞ 注:

- $\hat{y}$ 是 $\mathrm{E}(y)$ 的无偏估计.

$$\mathrm{E}(\hat{y}_i)=\mathrm{E}\left(\hat{\beta}_0+\hat{\beta}_1 x_i\right)=\beta_0+\beta_1 x_i=\mathrm{E}(y_i)$$

- $\bar{y}=\bar{\hat{y}}$
- 验证代码

```
mean(fire$loss) == mean(fitted(fire_lm))
```

☞ $\hat{\beta}_0, \hat{\beta}_1$ 的方差

$\hat{\beta}_0, \hat{\beta}_1$ 的方差

$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{L_{xx}}$, $\mathrm{Var}(\hat{\beta}_0) = \left[\frac{1}{n} + \frac{\bar{x}^2}{L_{xx}}\right]\sigma^2$, $\mathrm{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = -\frac{\bar{x}}{L_{xx}}\sigma^2$.

【证】因为 $\hat{\beta}_1 = \beta_1 + \frac{1}{L_{xx}}\sum_{i=1}^n d_i\varepsilon_i$，所以

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_1) &= \frac{1}{L_{xx}^2}\mathrm{Var}\left(\sum_{i=1}^n d_i\varepsilon_i\right) = \frac{1}{L_{xx}^2}\sum_{i=1}^n d_i^2\mathrm{Var}(\varepsilon_i) \\ &= \frac{1}{L_{xx}^2}\sum_{i=1}^n d_i^2\sigma^2 \\ &= \frac{\sigma^2}{L_{xx}}\end{aligned}$$

因为

$$\hat{\beta}_0 = \beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x} + \bar{\varepsilon}$$

所以

$$\begin{aligned}
\mathrm{Var}(\hat\beta_0) &= \mathrm{Var}\left[(\beta_1-\hat\beta_1)\bar{x}+\bar\varepsilon\right] \\
&= \bar{x}^2\mathrm{Var}(\hat\beta_1)+\frac{1}{n}\sigma^2-2\bar{x}\mathrm{Cov}(\beta_1-\hat\beta_1,\bar\varepsilon) \\
&= \left(\frac{\bar{x}^2}{L_{xx}}+\frac{1}{n}\right)\sigma^2-2\bar{x}\mathrm{Cov}\left(\frac{1}{L_{xx}}\sum_{i=1}^n d_i\varepsilon_i,\frac{1}{n}\sum_{i=1}^n\varepsilon_i\right) \\
&= \frac{\sigma^2\sum_{i=1}^n x_i^2}{nL_{xx}}+\mathbf{0} \\
&= \frac{\sigma^2\sum_{i=1}^n x_i^2}{nL_{xx}}
\end{aligned}$$

☞ **注:**

- $\mathrm{Var}(\hat\beta_1)$ 表示估计量 $\hat\beta_1$ 的稳定性, 该值越大表明估计量 $\hat\beta_1$ 的稳定性越差.
- $\mathrm{Var}(\hat\beta_1)$ 值取决于误差项方差 σ^2 和解释变量总的变异 L_{xx}, σ^2 越大 $\mathrm{Var}(\hat\beta_1)$ 越大, L_{xx} 越大 $\mathrm{Var}(\hat\beta_1)$ 越小.
- 以上分析表明, 在选择样本时, 应使解释变量 x 的观测值尽可能分散.
- $\hat\beta_1$ 的标准差为 $sd(\hat\beta_1)=\frac{\sigma}{\sqrt{L_{xx}}}$.

- $\hat{\beta}_0, \hat{\beta}_1$ 是关于正态随机变量 $y_1, y_2, \cdots, y_n$ 的线性函数，故 $\hat{\beta}_0, \hat{\beta}_1$ 都服从正态分布

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2 \sum_{i=1}^n x_i^2}{nL_{xx}}\right), \quad \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{L_{xx}}\right)$$

- 因 $\mathrm{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = -\frac{\bar{x}}{L_{xx}}\sigma^2$，可知当 $\bar{x} = 0$ 时 $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 不相关，在正态分布假设下独立；而 $\bar{x} \neq 0$ 时，不独立.
- 在 G-M 假设下，$\hat{\beta}_0$ 与 $\hat{\beta}_1$ 是 β_0 和 β_1 的最优线性无偏估计 (best linear unbiasd estimator, BLUE).
- 在给定点 x_i 处的拟合值

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \sim N\left(\beta_0 + \beta_1 x_i, \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{L_{xx}}\right)\sigma^2\right)$$

  由此可见，$\hat{y}_i$ 是 $\mathrm{E}(y_i|x_i)$ 的无偏估计，且其方差随着 x_i 与 $\bar{x}$ 距离 $|x_i - \bar{x}|$ 增大而增大，这说明在应用回归方程进行预测和控制时，给定的 x_i 的值离 $\bar{x}$ 越远，效果不理想的可能性越大.
- 取回归系数估计值的方差

```
vcov(fire_lm)
```

☞ t-检验

- t-检验的用途: 常用来检验回归系数的显著性 (检验解释变量 x 对被解释变量 y 的影响是否显著).
- 原假设和备择假设

$$H_0 : \beta_1 = 0, \quad H_1 : \beta \neq 0$$

- 检验统计量

$$T = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / L_{xx}}} = \frac{\hat{\beta}_1 \sqrt{L_{xx}}}{\hat{\sigma}} \sim t(n-2)$$

其中

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

且

$$\mathrm{E}(\hat{\sigma}^2) = \sigma^2, \quad \frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2), \quad \hat{\sigma} = \sqrt{\hat{\sigma}^2}$$

- 应用: 给定显著性水平 α, 双侧检验的临界值为 $t_{\frac{\alpha}{2}}(n-2)$. 当 $|T| \geqslant t_{\frac{\alpha}{2}}(n-2)$ 时拒绝原假设 H_0, 即认为 β_1 显著不等于零. 当 $|T| < t_{\frac{\alpha}{2}}(n-2)$ 时接受原假设 H_0, 即认为 $\beta_1 = 0$.

☞ *F*-检验

- 总平方和 (sum of squares for total, SST)

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

SST 度量了 y 中总样本变异. R 计算代码 `sst = sum((y - mean(y))^2)`

- 回归平方和 (sum of squares for regression, SSR)

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

SSR 度量了 $\hat{y}$ 中的总变异，或度量了解释变量所能够解释的 y 中的变异部分。

- 残差平方和 (sum of squares for errors, SSE)

$$SSE = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

SSE 度量了扰动项的变异，或度量了解释变量不能解释的 y 中的变异部分。
计算残差平方和 `deviance(fire_lm)`

- 总平方和，解释平方和，残差平方和存在如下关系：

$$SST = SSR + SSE$$

【证】

$$\begin{aligned} SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 \\ &= \sum_{i=1}^{n}[\hat{\varepsilon}_i + (\hat{y}_i - \bar{y})]^2 \\ &= \sum_{i=1}^{n}\hat{\varepsilon}_i^2 + 2\sum_{i=1}^{n}\hat{\varepsilon}_i(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\ &= SSE + SSR + 2\sum_{i=1}^{n}\hat{\varepsilon}_i(\hat{y}_i - \bar{y}) \\ &= SSR + SSE \end{aligned}$$

- F-检验的用途: 检验回归方程的整体显著性.
- 原假设与备择假设:

$$H_0 : \beta_1 = 0, \quad H_1 : \beta_1 \neq 0$$

- 检验统计量

$$F = \frac{SSR/1}{SSE/(n-2)} \sim F(1, n-2)$$

- 应用：如果回归方程是显著的，则 SSR 相对较大，即自变量 x 能够解释的 y 的波动部分较大，不能被自变量变化解释的部门 SSE 较小. 故给定显著性水平 α，如果 $F \geqslant F_\alpha(1, n-2)$，则表明回归方程是显著的，拒绝原假设. 否则接受原假设.
- 一元线性回归方差分析表

表: 一元线性回归方差分析表

方差来源	自由度	平方和	均方	F-值	p-值
回归	1	SSR	$SSR/1$	$\frac{SSR/1}{SSE/(n-2)}$	$P(F > F\_value) = p$
残差	$n-2$	SSE	$SSE/(n-2)$		
总和	$n-1$	SST			

- 方差分析 R 代码

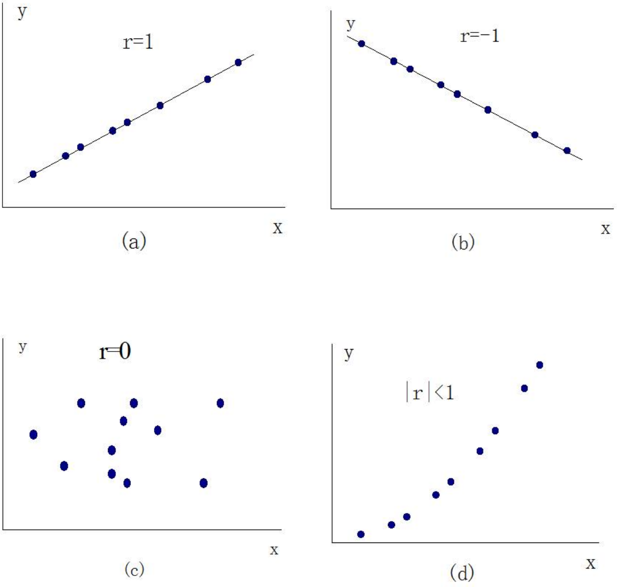
```
anova(fire_lm)                      # F 检验方差分析表
```

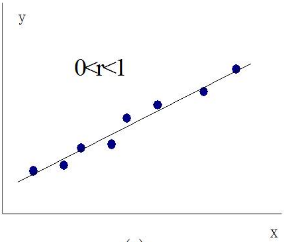
☞ 相关系数的显著性检验

- 可用变量间的简单相关系数检验回归方程的显著性.
- 数据 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ 的 Pearson 相关系数

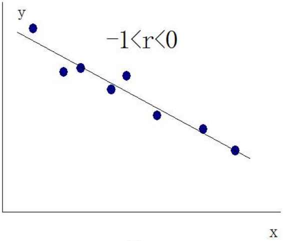
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}} \tag{15}$$

- $0 \leqslant |r| \leqslant 1$
- 总体相关系数 ρ
  - ✧ 高度相关: $|\rho| \geqslant 0.8$
  - ✧ 中度相关: $0.5 \leqslant |\rho| < 0.8$
  - ✧ 低度相关: $0.3 \leqslant |\rho| < 0.5$
  - ✧ 相关程度极弱: $0 < |\rho| < 0.3$
  - ✧ 不相关: $|\rho| = 0$
- 相关系数图例


y
r=1
x
(a)
y
r=−1
x
(b)
y
r=0
x
(c)
y
|r|<1
x
(d)



(e)



(f)

- 相关系数检验统计量. 因

$$r = \frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}} = \hat{\beta}_1 \sqrt{\frac{L_{xx}}{L_{yy}}} \tag{16}$$

检验统计量为

$$T = \frac{\sqrt{n-2}r}{\sqrt{1-r^2}}$$

- 相关系数显著性检验指令

```
cor.test(x,y)
```

- 注: 相关系数的显著性检验与相关程度强弱属于不同的概念.

☞ 三种检验的关系

- 对一元线性回归模型，回归系数的 t 检验、回归方程显著性的 F 检验、相关系数的显著性检验结果是完全一致的.
- 对多元线性回归模型，三种检验考虑的问题是不同的，属于三种不同的检验.

☞ 决定系数

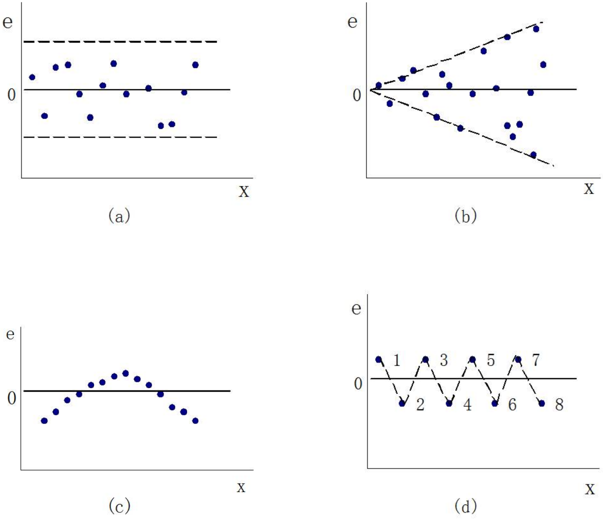
- 决定系数 (coefficient of determination)，又称为判定系数、确定系数，它是回归平方和与总离差平方和的比值，表示因变量的波动中可由自变量变化解释的比例.

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

- 注：
  - ✧ 决定系数 R^2 是反映回归模型与样本观测值拟合优度的相对指标，$R^2 \in [0, 1]$，R^2 越接近 1，表明模型拟合效果越好，R^2 越接近 0，表明模型拟合效果有待改进.
  - ✧ 样本容量 n 较小时，可能得到较大 R^2，一般属于虚假现象，不表示模型拟合效果好.
  - ✧ 即使在样本容量很大的情形得到较大的 R^2，也不能肯定自变量和因变量之间是线性关系.
  - ✧ 得到的 R^2 较小时，不代表模型一定不好.
- R 代码

```
summary(fire_lm)$r.square
```

☞ 残差图



☞ 残差的性质

① $E(e_i) = 0$

② $Var(e_i) = \left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{L_{xx}}\right]\sigma^2 = (1 - h_{ii})\sigma^2$

其中 $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{L_{xx}}$ 称为杠杆值, 且 $0 < h_{ii} < 1$. x_i 越接近 $\bar{x}$, h_{ii} 的值越小, 相应残差的方差越大. x_i 越是远离 $\bar{x}$, h_{ii} 的值越大, 相应残差的方差越小.

③ 残差满足 $\sum_{i=1}^{n} e_i = 0$, $\sum_{i=1}^{n} e_i x_i = 0$. 该性质表明残差 $e_1, e_2, \cdots, e_n$ 之间是相关的, 并不是独立的.

☞ 改进的残差

- 残差分析中, 一般认为残差绝对值超过 $2\hat{\sigma}$ 或 $3\hat{\sigma}$ 的观测点为异常值. 但直接依据残差很难判断.
- 标准化残差

$$ZRE_i = \frac{e_i}{\hat{\sigma}}$$

$|ZRE_i| > 3$ 的观测可直接判断为异常观测. 标准化残差没有解决残差方法不等的问题.

- 学生化残差

$$SRE_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

☞ 残差相关 R 代码

```
fire_resid <- residuals(fire_lm)        # 提取残差
sum(fire_resid)                         # 计算残差和，验证残差性质 3
crossprod(fire$distance,fire_resid)     # 验证残差性质 3
plot(fire_resid~fire$distance,pch=16,ylim=c(-5,5))
                                        # 绘制残差图
fire_length <-length(fire_resid)        # 计算残差数量
sigma <- summary(fire_lm)$sigma         # 提取估计的标准差
y <- numeric(fire_length)               # 绘制残差图中的三条参考线
y1 <- rep(2*sigma,fire_length)
y2 <- rep(-2*sigma,fire_length)
points(fire$distance,y,type="l",lty=1)
points(fire$distance,y1,type="l",lty=6)
points(fire$distance,y2,type="l",lty=6)

resid_stand <- fire_resid/sigma         # 计算标准化残差
resid_stu <- rstudent(fire_lm)          # 求学生化残差

which(abs(resid_stand)>3)               # 异常值检验
which(abs(resid_stu)>3)                 # 异常值检验
```

2.6 回归系数的区间估计

☞ **回归系数的区间估计**

- $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{L_{xx}}\right)$
- $T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2/L_{xx}}} \sim t(n-2)$
- 故 β_1 的置信度为 $1-\alpha$ 的置信区间为

$$\left(\hat{\beta}_1 - t_{\frac{\alpha}{2}}\frac{\hat{\sigma}}{\sqrt{L_{xx}}}, \hat{\beta}_1 + t_{\frac{\alpha}{2}}\frac{\hat{\sigma}}{\sqrt{L_{xx}}}\right)$$

☞ **R 代码**

```
fire_lm <- lm(loss~distance,data=fire) # 最小二乘回归估计

coefficients(fire_lm)                  # 取回归系数估计值
confint(fire_lm)                       # 取回归系数区间估计，缺省水平 95%
confint(fire_lm,parm=2)                # 取第2个系数的区间估计，缺省水平95%
confint(fire_lm,parm=2,level=0.99)     # 取第2个系数的区间估计，缺省水平99%
```

☞ 单值预测

- 设有估计的回归方程为 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. 现在给定自变量 x_0, 预测相应的因变量为

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

☞ 区间预测

问题: 对给定的自变量 x_0 和给定的显著性水平 α, 找一个区间 (T_1, T_2), 使得

$$\mathrm{P}(T_1 < y_0 < T_2) = 1 - \alpha$$

① 因变量新值的区间预测

✧ 首先计算估计值 $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

✧ 导出 $\hat{y}_0$ 的分布

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_0 = \sum_{i=1}^{n} \left[\frac{1}{n} + \frac{(x_i - \bar{x})(x_0 - \bar{x})}{L_{xx}} \right] y_i$$

故

$$\mathrm{Var}(\hat{y}_0) = \sum_{i=1}^{n} \left[\frac{1}{n} + \frac{(x_i - \bar{x})(x_0 - \bar{x})}{L_{xx}} \right]^2 \mathrm{Var}(y_i) = \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}} \right] \sigma^2$$

所以

$$\hat{y}_0 \sim N\left(\beta_0+\beta_1 x_0, \left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{L_{xx}}\right)\sigma^2\right)$$

记

$$h_{00}=\frac{1}{n}+\frac{(x_0-\bar{x})^2}{L_{xx}}$$

则

$$\hat{y}_0 \sim N\left(\beta_0+\beta_1 x_0, h_{00}\sigma^2\right)$$

预测新值时误差的方差为

$$\mathrm{Var}(y_0-\hat{y}_0)=\mathrm{Var}(y_0)+\mathrm{Var}(\hat{y}_0)=(1+h_{00})\sigma^2$$

而

$$\mathrm{E}(y_0-\hat{y}_0)=0$$

故有

$$y_0-\hat{y}_0 \sim N(0,(1+h_{00}\sigma^2))$$

进而有统计量

$$T=\frac{y_0-\hat{y}_0}{\sqrt{1+h_{00}}\hat{\sigma}} \sim t(n-2)$$

故 y_0 的置信度为 $1-\alpha$ 的置信区间为

$$\left(\hat{y}_0-t_{\frac{\alpha}{2}}(n-2)\sqrt{1+h_{00}}\hat{\sigma}, \hat{y}_0+t_{\frac{\alpha}{2}}(n-2)\sqrt{1+h_{00}}\hat{\sigma}\right)$$

② 因变量新值平均值的区间预测

因为

$$\hat{y}_0 \sim N\left(\beta_0 + \beta_1 x_0, h_{00}\sigma^2\right)$$

故

$$\hat{y}_0 - \mathrm{E}(y_0) \sim N\left(0, h_{00}\sigma^2\right)$$

进而可得置信水平为 $1-\alpha$ 的置信区间为

$$(\hat{y}_0 - t_{\frac{\alpha}{2}}(n-2)\sqrt{h_{00}}\hat{\sigma}, \hat{y}_0 + t_{\frac{\alpha}{2}}(n-2)\sqrt{h_{00}}\hat{\sigma})$$

③ R 代码

```
newdata <- data.frame(distance=2.8)
predict(fire_lm, newdata)                    # 单值预测
predict(fire_lm, newdata, interval="confidence")
                                             # 均值的区间预测，置信水平为 95%
predict(fire_lm, newdata, interval="confidence", level=0.99)
                                             # 均值的区间预测，置信水平为 99%
predict(fire_lm, newdata, interval="prediction")
                                             # 单值的区间预测，置信水平为 95%
predict(fire_lm, newdata, interval="prediction", level=0.99)
                                             # 单值的区间预测，置信水平为 99%
```

- R 代码: 将数据点, 预测估计曲线, 预测区间估计曲线, 置信区间曲线画在同一张图上.

```
M <- max(fire$distance)
m <- min(fire$distance)
newdata <- data.frame(distance=seq(m,M,by=0.01))

pp <- predict(fire_lm, newdata, interval="prediction")
pc <- predict(fire_lm, newdata, interval="confidence")
par(mai=c(0.8,0.8,0.2,0.2))
matplot(newdata$distance, cbind(pp,pc[,-1]),type="l",
        xlab="distance", ylab="loss", lty=c(1,5,5,2,2),
        col=c("blue","red","red","brown","brown"),
        lwd=2)
with(points(distance,loss,cex=1.4,pch=21,col="red",bg="orange"),data=fire)
legend(0.8, 45,
       c("Points","Fitted", "Prediction","Confidence"),
       pch = c(19,NA,NA,NA), lty=c(NA,1,5,2),
       col = c("orange","blue","red","brown"))
```

☞ 控制问题

- 问题: 为了以概率 $1-\alpha$ 将因变量 y 控制在区间 (T_1, T_2) 内, 如何控制自变量 x?
- 分析: 问题等价于

$$P(T_1 < y < T_2) = 1 - \alpha$$

即

$$\begin{cases} \hat{y}(x) - t_{\frac{\alpha}{2}}(n-2)\sqrt{1+h_{00}}\hat{\sigma} > T_1 \\ \hat{y}(x) + t_{\frac{\alpha}{2}}(n-2)\sqrt{1+h_{00}}\hat{\sigma} < T_2 \end{cases}$$

将 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ 代入上式, 则有

① 当 $\hat{\beta}_1 > 0$ 时,

$$\frac{T_1 + t_{\frac{\alpha}{2}}(n-2)\sqrt{1+h_{00}}\hat{\sigma} - \hat{\beta}_0}{\hat{\beta}_1} < x < \frac{T_2 - t_{\frac{\alpha}{2}}(n-2)\sqrt{1+h_{00}}\hat{\sigma} - \hat{\beta}_0}{\hat{\beta}_1}$$

② 当 $\hat{\beta}_1 < 0$ 时,

$$\frac{T_2 - t_{\frac{\alpha}{2}}(n-2)\sqrt{1+h_{00}}\hat{\sigma} - \hat{\beta}_0}{\hat{\beta}_1} < x < \frac{T_1 + t_{\frac{\alpha}{2}}(n-2)\sqrt{1+h_{00}}\hat{\sigma} - \hat{\beta}_0}{\hat{\beta}_1}$$

作业

☞ **P53**

2.1
2.2
2.3
2.4
2.5
2.6
2.7
2.8
2.9
2.10
2.11
2.12
2.13