第 4 章　违背基本假设的情况

李　杰

数据科学学院，浙江财经大学

2017 年 11 月 14 日

内容提要

☞ **G-M 条件**

$$\begin{cases} \mathrm{E}(\varepsilon_i) = 0, & i = 1, 2, \cdots, n \\ \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2, & i = j \\ 0, & i \neq j \end{cases} & i, j = 1, 2, \cdots, n \end{cases}$$

☞ **违背的基本假设**

① **异方差问题:**

$$\mathrm{Var}(\varepsilon_i) \neq \mathrm{Var}(\varepsilon_j) \quad (i \neq j)$$

② **自相关问题:**

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) \neq 0, \quad (i \neq j)$$

4.1 异方差性产生的背景和原因

☞ **异方差产生的原因**

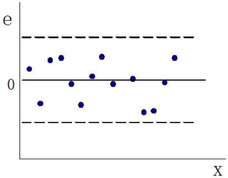
- 实际问题错综复杂.

☞ **异方差性带来的问题:**

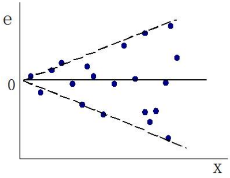
- 参数的最小二乘估计是无偏估计、相合估计, 但不是有效估计.
- 参数的显著性检验失效.
- 回归方程的应用效果不理想.

☞ 异方差的检验

① 残差图分析法



(a)



(b)

② 等级相关系数法

1. 做 y 关于 x 的普通最小二乘回归，求出残差 e_i;
2. 计算 $|e_i|$，将 x_i 或 $|e_i|$ 按递增或递减次序排列后分成等级，并计算等级相关系数

$$r_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} d_i^2$$

其中 n 为样本容量，d_i 为对应于 x_i 和 $|e_i|$ 的等级差数.

3. 做等级相关系数的显著性检验. 在 $n > 8$ 的条件下，检验统计量为

$$t = \frac{\sqrt{n-2}\,r_s}{\sqrt{1-r_s^2}} \sim t(n-2)$$

☞ 加权最小二乘估计;

- 目标函数

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

- 加权最小二乘估计量

$$\hat{\beta}_{0\omega} = \bar{y}_\omega - \hat{\beta}_{1\omega} \bar{x}_\omega$$

$$\hat{\beta}_{1\omega} = \frac{\sum\limits_{i=1}^{n} \omega_i (x_i - \bar{x}_\omega)(y_i - \bar{y}_\omega)}{\sum\limits_{i=1}^{n} \omega_i (x_i - \bar{x}_\omega)^2}$$

其中

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

多元加权最小二乘法

- **多元线性回归模型**

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad (i = 1, 2, \cdots, n)$$

- **加权离差平方和**

  当误差项 ε_i 存在异方差时, 加权离差平方和为

$$Q_\omega(\beta_0, \beta_1, \cdots, \beta_p) = \sum_{i=1}^{n} \omega_i (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip})^2$$

- **加权矩阵与加权估计量**

  设有加权矩阵

$$\Omega = \begin{bmatrix} \omega_1 & & & \\ & \omega_2 & & \\ & & \ddots & \\ & & & \omega_n \end{bmatrix}$$

  加权最小二乘估计为

$$\hat{\beta}_\omega = (\mathbf{X}'\Omega\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{y}$$

☞ 对异方差的检验

- **模型:** 考虑一般的多元线性回归模型

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon \tag{1}$$

- **问题:** 检验"同方差"假设是否成立. 对应的原假设

$$H_0 : \mathrm{Var}(\varepsilon | x_1, \cdots, x_k) = \sigma^2$$

因 $\mathrm{E}(\varepsilon | x_1, \cdots, x_k) = 0$, 故同方差的虚拟假设等价于

$$H_0 : \mathrm{E}(\varepsilon^2 | x_1, \cdots, x_k) = \mathrm{E}(\varepsilon^2) = \sigma^2$$

- **分析:** 如果 H_0 不成立, 则 ε^2 可能是关于 $x_1, \cdots, x_k$ 的函数, 一个简单方法是做回归

$$\varepsilon^2 = \delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k + u \tag{2}$$

然后用 F 统计量或 LM 统计量检验假设

$$\delta_1 = 0, \delta_2 = 0, \cdots, \delta_k = 0$$

- **应用方法:** 因 ε_i 不可观测, 但可用 OLS 残差 $\hat{\varepsilon}_i$ 作为其估计值, 故可估计方程

$$\hat{\varepsilon}^2 = \delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k + \nu \tag{3}$$

记回归 (3) 的 R^2 为 $R^2_{\hat{\nu}}$, 则检验 $x_1, \cdots, x_k$ 联合显著性的 F 统计量为

$$F = \frac{R^2_{\hat{\nu}}/k}{(1 - R^2_{\hat{\nu}})/(n-k-1)} \overset{d}{\sim} F_{k,n-k-1} \tag{4}$$

LM 统计量为

$$LM = nR^2_{\hat{\nu}} \overset{d}{\sim} \chi^2_k \tag{5}$$

☞ **布罗施-帕甘异方差检验** (Breusch-Pagan test for heteroskedasticity, BP test)

- **检验步骤:**
  ① 用 OLS 方法估计多元回归模型 (1), 得到 OLS 残差 $\hat{\varepsilon}_i$;
  ② 计算残差平方值 $\hat{\varepsilon}_i^2$, 并作回归 (3), 得到回归的 R^2, 记为 $R^2_{\hat{\nu}}$;
  ③ 利用 $R^2_{\hat{\nu}}$ 计算检验联合显著性的 F 统计量或 LM 统计量, 并依据 F 分布或 χ^2 分布计算 p 值. 如果该 p 值小于给定的显著性水平, 则拒绝同方差的原假设.
- **注解:**
  ① BP 检验拒绝原假设后, 则须使用异方差稳健方法对原回归的标准误、统计量进行调整. 或者使用加权最小二乘法.

② 如果判断异方差只取决于部分解释变量, 则需对 BP 检验做释放的修改: 做 $\hat{\varepsilon}_i^2$ 对这部分解释变量的回归.

- BP 检验的 R 指令:
  1. car 包中的函数 ncv.test();
  2. lmtest 包中的函数 bptest()
  3. 示例

```
library(lmtest)

sav<-read.dta("E:/statafiles/SAVING.dta", convert.factors = FALSE)

lmsav <- lm(sav~inc+size+educ+age+black,data=sav)

bptest(lmsav)

bptest(lmsav,studentize=FALSE)    # 去掉 Koenker 的学生化方法,
                                  # 该检验结果与 ncv.test() 一致

install.packages(car)             # 另一种检验方法
library(car)

ncvTest(lmsav)
```

☞ White 异方差检验

- **弱化的同方差假设:** 同方差假设 $\mathrm{Var}(\varepsilon|x_1, x_2, \cdots, x_k) = \mathrm{E}(\varepsilon^2) = \sigma^2$ 可以弱化为“扰动项的平方与所有的解释变量 (x_i)、解释变量的平方项 (x_i^2)、所有解释变量的交叉乘积 ($x_i x_j$, $i \neq j$) 都不相关”.
- **White 检验:** 检验 OLS 残差平方和关于所有解释变量、所有解释变量的平方、所有解释变量的交叉项的回归是否显著. 譬如, 假定原模型为

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

OLS 残差平方和为 $\hat{\varepsilon}_i^2$. **White 异方差检验** (White test for heteroskedasticity) 就是检验回归

$$\hat{\varepsilon}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + \delta_4 x_1^2 + \delta_5 x_2^2 + \delta_6 x_3^2 + \delta_7 x_1 x_2 + \delta_8 x_1 x_3 + \delta_9 x_2 x_3 + \nu$$

的模型显著性 (可用 LM 统计量或 F 统计量).
- **White 检验的缺陷:** 辅助回归中变量过多, 占用太多的自由度, 当解释变量的个数为 k 时, 与 BP 检验相比, White 检验要多占用 $k(k+1)/2$ 个自由度.
- **改进的 White 检验:**
  ① 用 OLS 法估计模型 (1), 得到 OLS 残差 $\hat{\varepsilon}$ 和拟合值 $\hat{y}$.

② 计算参差平方 $\hat{\varepsilon}^2$ 以及拟合值的平方 $\hat{y}^2$, 并作回归

$$\hat{\varepsilon}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + \nu$$

保留其 R^2, 记为 $R^2_{\hat{\varepsilon}}$.

③ 利用 $R^2_{\hat{\varepsilon}}$ 构造 F 统计量或 LM 统计量, 并计算 p 值. 最后比较该 p 值与给定的显著性水平, 判断是否拒绝原假设.

☞ **加权最小二乘估计** (weighted least squared method, WLS)

- **样本回归模型:** 样本回归模型为

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i \tag{6}$$

记 $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ik})$.

☞ **假设:** 如果数据存在异方差问题, 则可假定

$$\mathrm{Var}(\varepsilon|\mathbf{x}) = \sigma^2 h(\mathbf{x}) \tag{7}$$

其中 $h(\mathbf{x}) > 0$. 对于第 i 个观测, 其方差 σ_i^2 具有表达式 $\sigma_i^2 = \sigma^2 h(\mathbf{x}_i) = \sigma^2 h_i$.

- **加权最小二乘估计步骤:**

① 在模型 (6) 等号两侧同事除以 $\sqrt{h_i}$, 得到模型

$$\frac{y_i}{\sqrt{h_i}} = \beta_0 \frac{1}{\sqrt{h_i}} + \beta_1 \frac{x_{i1}}{\sqrt{h_i}} + \cdots + \beta_k \frac{x_{ik}}{\sqrt{h_i}} + \frac{\varepsilon_i}{\sqrt{h_i}}$$

整理该模型可得

$$y_i^* = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + \cdots + \beta_k x_{ik}^* + \varepsilon_i^* \tag{8}$$

因 $\mathrm{Var}(\varepsilon_i^*) = (\sigma^2 h_i)/h_i = \sigma^2$, 故模型 (8) 满足 G-M 条件.

② 用 OLS 估计模型 (8) 得到的参数估计量称为**广义最小二乘估计** (generalized least squared estimators, GLS).

③ 估计模型 (8) 得到的标准误、t 统计量、F 统计量满足 Gauss-Markov 定理.

④ 模型 (8) 残差平方和除以自由度就是 σ^2 的无偏一致估计.

⑤ 当 $h_i = \mathrm{Var}(\varepsilon_i|\mathbf{x}_i)$ 时，GLS 估计是有效估计.

- **注解:** 当权函数 $h(\mathbf{x})$ 选择不正确时, WLS 的性质.

① WLS 与 OLS 一样, 是无偏且一致估计.

② t 统计量和 F 统计量都不正确.

③ 在强异方差情形, 即使 WLS 使用了错误的加权函数, 估计量的性质也由于完全忽略异方差的 OLS 估计.

☞ 可行广义最小二成估计

- **问题:** 在多数情形, 模型存在异方差时, 异方差的准确形式未知, 即权函数 $h(\mathbf{x}_i)$ 未知.

- **方法:** 模型化 $h(\mathbf{x})$, 并用数据估计其中的未知参数, 从而得到每个 h_i 的估计值 $\hat{h}_i$, 并在 GLS 中用 $\hat{h}_i$ 代替 h_i, 该方法称为**可行广义最小二乘估计** (feasible generalized least squared estimators, FGLS). 譬如

$$\mathrm{Var}(\varepsilon|\mathbf{x}) = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k) \tag{9}$$

其中, $h(\mathbf{x}) = \exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k)$ (注: 与线性函数相比, 指数函数形式的 h 可确保 $h(\mathbf{x}) > 0$).

- **推导:** 对于假定 (6), 为估计其中的参数 $\delta_0, \delta_1, \cdots, \delta_k$, 令

$$\varepsilon^2 = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k)\nu$$

其中 $\mathrm{E}(\nu|\mathbf{x}) = 1$. 两边取对数可得

$$\log(\varepsilon^2) = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k + e \tag{10}$$

其中 $\alpha_0 = \log(\sigma^2) + \delta_0$, $e = \log(\nu)$, 且 $\mathrm{E}(e|\mathbf{x}) = 0$.

☞ 纠正异方差的一个可行 GLS 法

① 对模型 (1), 作 y 对 $x_1, x_2, \cdots, x_k$ 的 OLS 回归并得到残差 $\hat{\varepsilon}$;
② 计算残差平方值 $\hat{\varepsilon}^2$, 再计算其对数值 $\log(\hat{\varepsilon}^2)$;

③ 作回归

$$\log(\hat{\varepsilon}^2) = \alpha_0 + \delta_1 x_1 + \cdots + \delta_k x_k + e \tag{11}$$

并得到拟合值 $\hat{g}_i = \hat{\alpha}_0 + \hat{\delta}_1 x_{i1} + \cdots + \hat{\delta}_k x_{ik}$.

④ 计算拟合值的指数：$\hat{h}_i = \exp(\hat{g}_i)$

⑤ 以 $1/\hat{h}_i$ 为权重，用 WLS 估计原模型.

注解：

① FGLS 估计量虽不是无偏估计，但是一致估计，且比 OLS 更有效；

① 辅助回归 (10) 可替换成

$$\log(\hat{\varepsilon}^2) = \eta_0 + \eta_1 \hat{y} + \eta_2 \hat{y}^2 + \nu \tag{12}$$

☞ 案例：香烟需求模型

- 利用数据集 SMOKE.DTA 估计烟民对香烟的日需求量。用 OLS 估计的模型为

$$\begin{aligned}
\widehat{cigs} = & -\underset{(24.08)}{3.64} + \underset{(0.728)}{0.88}\log(income) - \underset{(5.773)}{0.751}\log(cigpric) - \underset{(0.167)}{0.501}educ \\
& + \underset{(0.160)}{0.771}age - \underset{(0.0017)}{0.009}age^2 - \underset{(1.11)}{2.83}restaurn \\
& n = 807, \quad R^2 = 0.0526
\end{aligned} \tag{13}$$

其中 $cigs$ 表示日吸烟量 (单位：支)，$income$ 表示烟民的年收入，$cigpric$ 表示每包香烟的价格，$educ$ 表示烟民的受教育程度，age 表示烟民的年龄，二值变量 $restaurn = 1$ 表示烟民所在州在餐馆禁烟。

- 对该模型进行 BP 异方差检验，可发现 $LM = 32.28$，是数据异方差存在的极强证据。
- 用可行 GLS 重新估计模型，得到

$$\begin{aligned}
\widehat{cigs} = & \underset{(17.80)}{5.64} + \underset{(0.44)}{1.301}\log(income) - \underset{(4.46)}{2.941}\log(cigpric) - \underset{(0.120)}{0.463}educ \\
& + \underset{(0.097)}{0.482}age - \underset{(0.0009)}{0.0056}age^2 - \underset{(0.80)}{3.46}restaurn \\
& n = 807, \quad R^2 = 0.1134
\end{aligned} \tag{14}$$

- **注解：**
  1. 如果 OLS 估计和 GLS 估计存在符号不同却都是统计显著的，或者估计值差异较大且都是显著的，则说明模型可能不满足 Gauss-Markov 条件。
  2. 在本例中，OLS 估计与 GLS 估计符号都相同，估计值差异较大的统计量都是统计上不显著的。故影响不是太大。
- **代码：**

```
library(foreign)
smoke <- read.dta("E:/statafiles/SMOKE.DTA", convert.factors = FALSE)

# OLS 估计
cig_lm <- lm(cigs~log(income)+log(cigpric)+educ+age+I(age^2)+restaurn,
                data=smoke)
summary(cig_lm)

# 辅助回归
logresidsq = log(resid(cig_lm)^2) # 求残差的平方
auxlm <- lm(logresidsq~log(smoke$income)+log(smoke$cigpric)+smoke$educ
            +smoke$age+I(smoke$age^2)+smoke$restaurn)
hath <- exp(fitted(auxlm))        # 求权函数 h
```

```
smoke$wcigs <- smoke$cigs /hath   # 数据变换
smoke$wloginc <- log(smoke$income)/hath
smoke$wlogcigp <- log(smoke$cigpric)/hath
smoke$weduc <- smoke$educ/hath
smoke$wage <- smoke$age/hath
smoke$wagesq <- smoke$age^2 /hath
smoke$wrest <- smoke$restaurn/hath
                                  # 用变换后的数据做加权回归
cig_wlm <- lm(wcigs~wloginc + wlogcigp + weduc + wage +  wagesq + wrest,
                data=smoke)
summary(cig_wlm)
```

☞ **自相关**

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) \neq 0 \quad (i \neq j)$$

☞ **自相关产生的背景和原因**

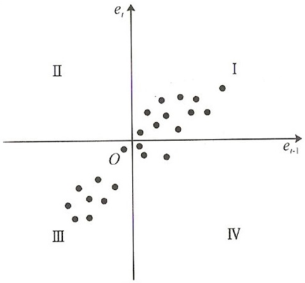
- 遗漏了关键变量.
- 经济变量的滞后性.
- 错误的模型设定.
- 蛛网现象.
- 因对数据加工整理导致的误差项之间产生的自相关.

☞ **自相关带来的问题**

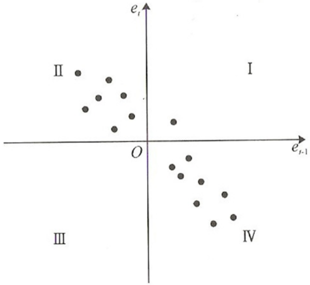
- 参数的最小二乘估计量不再是最小方差无偏估计;
- 均方误差 (MSE) 可能严重低估误差项的方差;
- 导致 t 检验和 F 检验的失效;
- 参数的最小二乘估计量仍是无偏、相合估计, 但自相关可能导致最小二乘估计量对波动异常敏感.
- 导致结构分析和预测的较大方差和错误解释.

☞ 自相关的诊断

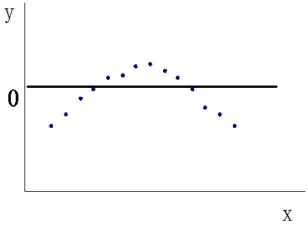
- 图示检验法



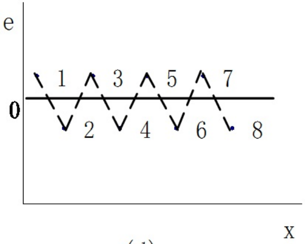
(a)



(b)



(c)



(d)

- **DW 检验**
  - **自相关系数**
    - 自相关系数: $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ 的自相关系数定义为

$$\rho = \frac{\sum_{t=2}^{n} \varepsilon_t \varepsilon_{t-1}}{\sqrt{\sum_{t=2}^{n} \varepsilon_t^2} \sqrt{\sum_{t=1}^{n-1} \varepsilon_{t-1}^2}}$$

    - $\rho \in [-1, 1]$
    - ρ 的估计量为

$$\hat{\rho} = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sqrt{\sum_{t=2}^{n} e_t^2} \sqrt{\sum_{t=1}^{n-1} e_{t-1}^2}}$$

  - **DW 检验**
    - 前提假设: 随机扰动项存在一阶自相关.

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t$$

✧ 检验问题对应的原假设

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

✧ 检验统计量

$$DW = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=2}^{n} e_t^2} = \frac{\sum_{t=2}^{n} e_t^2 + \sum_{t=2}^{n} e_{t-1}^2 - 2\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e_t^2}$$

$$= 2(1 - \hat{\rho})$$

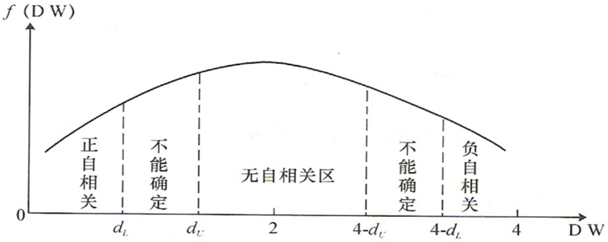
$$DW \in [0, 4]$$

✧ DW 的值与 $\hat{\rho}$ 的值之间的关系

$\hat{\rho}$	DW	误差项的自相关性
-1	4	完全负自相关
$(-1, 0)$	$(2, 4)$	负自相关
0	2	无自相关
$(0, 1)$	$(0, 2)$	正自相关
1	0	完全正自相关

✧ DW 值的应用

$0 \leqslant DW \leqslant d_L$	误差项 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ 间存在正自相关
$d_L \leqslant DW \leqslant d_U$	不能判断
$d_U \leqslant DW \leqslant 4 - d_U$	无自相关系
$4 - d_U \leqslant DW \leqslant 4 - d_L$	不能判断
$0 \leqslant DW \leqslant d_L$	负自相关



- *DW* 检验的局限性
  - ✧ *DW* 检验有两个不能确定的区域, 如果 *DW* 的值落入这两个区域, 则无法判断, 需要增大样本, 或者选取其他方法.
  - ✧ *DW* 检验统计量的上、下界表要求 $n > 15$, 如果样本容量过小, 则 *DW* 检验难做出正确的判断.
  - ✧ *DW* 检验不适合具有高阶相关的随机扰动项.

☞ 自相关问题的处理方法

- **迭代法**
  - ✧ 假设

$$
\begin{aligned}
& y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \\
& \varepsilon_t = \rho\varepsilon_{t-1} + u_t \\
& \begin{cases} \mathrm{E}(u_t) = 0, & t = 1, 2, \cdots, n \\ \mathrm{Cov}(u_t, u_s) = \begin{cases} \sigma^2, & t = s \\ 0, & t \neq s \end{cases} & t, s = 1, 2, \cdots, n \end{cases}
\end{aligned} \tag{15}
$$

  - ✧ 滞后一期模型

$$
y_{t-1} = \beta_0 + \beta_1 x_{t-1} + \varepsilon_{t-1} \tag{16}
$$

  - ✧ 准差分变换: (15) − (16) × ρ 可得

$$
y_t - \rho y_{t-1} = \beta_0(1-\rho) + \beta_1(x_t - \rho x_{t-1}) + (\varepsilon_t - \rho\varepsilon_{t-1})
$$

令 $y'_t = y_t - \rho y_{t-1}$，$x'_t = x_t - \rho x_{t-1}$，$u_t = \varepsilon_t - \rho\varepsilon_{t-1}$，$\beta'_0 = \beta_0(1-\rho)$, 则有

$$
y'_t = \beta'_0 + \beta_1 x'_1 + u_t \tag{17}
$$

✧ 注:

① 因 ρ 未知, 所以可用

$$\hat{\rho} \approx 1 - \frac{1}{2}DW$$

计算 ρ 的估计值, 如果式 (15) 中的误差项确实存在一阶自相关, 可式 (17) 已消除自相关关系.

② 式 (17) 中的 u_t 不一定满足 G-M 假设, 还需要对 u_t 进行 DW 检验.

③ 如果 u_t 存在自相关, 则需用迭代法消除一阶自相关关系.

- **差分法**

✧ 如果扰动项的一阶自相关程度较高时, 可用差分法.

✧ 差分法: 式 (15) – (16) 可得

$$y_t - y_{t-1} = \beta_1(x_t - x_{t-1}) + (\varepsilon_t - \varepsilon_{t-1})$$

令 $\Delta y_t = y_t - y_{t-1}$, $\Delta x_t = x_t - x_{t-1}$, $u_t = \varepsilon_t - \varepsilon_{t-1}$, 则有

$$\Delta y_t = \beta_1 \Delta x_t + u_t \tag{18}$$

✧ 用最小二乘法估计无解决回归模型 (18), 可得

$$\hat{\beta}_1 = \frac{\sum_{t=2}^{n} \Delta y_t \Delta x_t}{\sum_{t=2}^{n} \Delta x_t^2}$$

☞ R 代码: 迭代法

```
library(car)
library(timeSeries)

expend_lm <- lm(expend~income, data=expend)

my_dw <- dwtest(expend_lm) # lmtest 包中的 d-w 检验函数

rho <- 1-my_dw$dw/2

ts_expend <- as.timeSeries(expend$expend)     # 迭代法
ts_income <- as.timeSeries(expend$income)

i_expend <- na.omit(ts_expend-rho*lag(ts_expend,1))
i_income <- na.omit(ts_income-rho*lag(ts_income,1))

i_expend_lm <- lm(i_expend~i_income)
summary(i_expend_lm)
i_my_dw <- durbinWatsonTest(i_expend_lm) # car 包中的d-w检验函数
```

☞ **R 代码: 差分法**

```
dif_expend <- na.omit(diff(ts_expend))        # 差分法
dif_income <- na.omit(diff(ts_income))
dif_expend_lm <- lm(dif_expend~dif_income-1)
dwtest(dif_expend_lm)
```

4.5 BOX-COX 变换

- **背景:** 该方法由 Box 和 Cox 于 1964 年提出.
- **作用:** BOX-COX 变换可以消除异方差、自相关、误差非正态、回归函数非线性等问题.
- **BOX-COX 变换:** 如果所有因变量 y 的值都大于 0, 则做变换

$$y^{(\lambda)} = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln y, & \lambda = 0 \end{cases}$$

其中 λ 是待定参数. 如果存在因变量的值小于 0, 则做变换

$$y^{(\lambda)} = \begin{cases} \dfrac{(y+a)^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln(y+a), & \lambda = 0 \end{cases}$$

即先对 y 向右平移 a 个单位, 使得所有的因变量都大于 0 后再做 BOX-COX 变换.

- **注意:** 对于不同的 λ, 所做的变换也不同. 譬如
  - $\lambda = 0$ 时对应对数变换.
  - $\lambda = \frac{1}{2}$ 时对应平方根变换.
  - $\lambda = -1$ 时对应倒数变换.

☞ 主要方法

- 目的: 寻找合适的 λ, 使得变换后有

$$\mathbf{y}^{(\lambda)} = \left(y_1^{(\lambda)},\ y_2^{(\lambda)},\ \cdots\ y_n^{(\lambda)}\right)' \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

  从而满足线性模型的各项假设.
- 首先, 将 λ 当做常数, 求出 $\hat{\sigma}_\lambda^2$.
- 构造关于 λ 的似然函数

$$L_{\max}(\lambda) = \left(2\pi e\hat{\sigma}_\lambda^2\right)^{-\frac{n}{2}} |\mathbf{J}|$$

  其中 $\hat{\sigma}_\lambda^2 = \dfrac{1}{n}SSE(\lambda, y^{(\lambda)})$, $|\mathbf{J}| = \prod_{i=1}^n \left|\dfrac{dy_i^{(\lambda)}}{dy_i}\right| = \prod_{i=1}^n y_i^{(\lambda)}$.
- 令 $z^{(\lambda)} = \dfrac{y^{(\lambda)}}{|\mathbf{J}|}$, 对 $L_{\max}(\lambda)$ 取对数并略去与 λ 无关的常数项, 可得

$$\ln L_{\max}(\lambda) = -\frac{n}{2}\ln SSE(\lambda, z^{(\lambda)})$$

- 找出使得 $L_{\max}(\lambda)$ 最大, 即使得 $SSE(\lambda, z^{(\lambda)})$ 最小的 λ. 通常需要数值解法.

☞ 消除异方差

```
library(foreign)
library(MASS)
                                                        # 消除异方差
peking <- read.spss("D:\\documents\\MyDoc\\JobInZufe\\Courseware
         \\应用回归分析\\数据\\例3.2 北京市开发区.sav",to.data.frame =T)

peking_lm <- lm(Y~X1+X2, data=peking)
summary(peking_lm)

op <- par(mfrow=c(2,2), mar=.4+c(4,4,1,1),oma=c(0,0,2,0))
plot(fitted(peking_lm),resid(peking_lm), cex=1.2,pch=21,
           col="red",bg="orange",xlab="拟合值", ylab="残差")

boxcox(peking_lm, lambda=seq(0,1,by=0.1))
```

从第 2 张图大致可看出 $\lambda \approx 0.47$ 时, 似然函数值最大. 故确定 $\lambda = 0.47$.

```
lambda <- 0.47
Ylam <- (peking$Y^lambda-1)/lambda

lam_peking_lm <- lm(Ylam~peking$X1+peking$X2)
summary(lam_peking_lm)

plot(fitted(lam_peking_lm),resid(lam_peking_lm),cex=1.2,pch=21,
                 col="red",bg="orange", xlab="拟合值", ylab="残差")

library(car)                  # 再做异方差检验
ncvTest(lam_peking_lm)
```

估计的方程为

$$\hat{Y} = (4.196 + 0.024 * X1 + 0.006 * X2)^{\frac{1}{0.47}}$$

☞ 消除自相关

```
library(foreign)
library(MASS)                                    # 消除自相关

expend <- read.csv("D:\\documents\\MyDoc\\JobInZufe\\Courseware
                        \\应用回归分析\\数据\\expend_income.csv")

expend_lm <- lm(expend~income, data=expend)
summary(expend_lm)

op <- par(mfrow=c(2,2), mar=.4+c(4,4,1,1),oma=c(0,0,2,0))

plot(fitted(expend_lm),resid(expend_lm),cex=1.2,pch=21,col="red",
                           bg="orange", xlab="拟合值", ylab="残差")

boxcox(expend_lm, lambda=seq(-1,2,by=0.1))
```

从第 2 张图大致可看出 $\lambda \approx 1.15$ 时, 似然函数值最大. 故确定 $\lambda = 1.17$.

```
lambda <- 1.15
Ylam <- (expend$expend^lambda-1)/lambda

lam_expend_lm <- lm(Ylam~expend$income)
summary(lam_expend_lm)

plot(fitted(lam_expend_lm),resid(lam_expend_lm),cex=1.2,pch=21,
                col="red",bg="orange", xlab="拟合值", ylab="残差")

library(lmtest)              # 再做自相关检验
dwtest(lam_expend_lm)
```

估计的方程为

$$\widehat{expend} = (-823.079 + 2.966 * income)^{\frac{1}{1.15}}$$

关于因变量 y 的异常值

- **普通残差分析:** 对于个体 i, 如果其最小二乘残差 $|e_i| \geqslant 3\hat{\sigma}$, 则认为该个体是异常值.
- **标准化残差:** 如果标准化残差 $ZRE_i = \dfrac{e_i}{\hat{\sigma}}$ 的绝对值大于 3, 则认为该个体为异常值.
- **学生化残差:** 如果学生化残差 $SRE_i = \dfrac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$ 的绝对值大于 3, 则认为该个体为异常值, 其中 h_{ii} 为帽子矩阵 $H = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ 的主对角线元素.
- **删除残差:** 从样本中删除第 i 个个体后估计模型, 再求出第 i 个观测的残差 $e_{(i)}$. 可以证明 $e_{(i)} = \dfrac{e_i}{1-h_{ii}}$.
- **删除学生残差**

$$SRE_{(i)} = SRE_i \left(\frac{n-p-2}{n-p-1-SRE_i^2} \right)^{\frac{1}{2}}$$

☞ 关于自变量 x 的异常值

- **强影响点:** 因为 $\mathrm{Var}(e_i) = (1 - h_{ii})\sigma^2$, h_{ii} 调节 e_i 方差的大小, 故称 h_{ii} 为第 i 个观测的**杠杆值**. 杠杆值偏大的样本点称之为**强影响点**.
- **强影响点与异常点:** 强影响点不一定是异常点, 但是对回归方程影响很大.
- **库克距离**

$$D_i = \frac{e_i^2}{(p+1)\hat{\sigma}^2} \frac{h_{ii}}{(1-h_{ii})^2}$$

  库克距离反映了残差值 h_{ii} 和残差 e_i 的综合效应.
- **强影响点的判断:**
  - ✧ 因为 $tr(H) = \sum_{i=1}^{n} h_{ii} = p + 1$, 故杠杆值的均值为

$$\bar{h} = \frac{1}{n} \sum_{i=1}^{n} h_{ii} = \frac{p+1}{n}$$

    如果个体 i 的杠杆值 h_{ii} 超过 $\bar{h}$ 的 2 倍或 3倍, 则可认为个体 i 是强影响点.
  - ✧ 中心化杠杆值 $ch_i = h_{ii} - \frac{1}{n}$, 故 $\sum_{i=1}^{n} ch_i = p$, 中心化杠杆值的均值为 $\bar{ch} = \frac{p}{n}$. 如果个体 i 的中心化杠杆值 ch_{ii} 超过 $\bar{ch}$ 的 2 倍或 3倍, 则可认为个体 i 是强影响点.
  - ✧ $D_i < 0.5$ 时可认为个体 i 不是强影响点, 而 $D_i > 1$ 是可认为个体 i 是强影响点.

☞ R 代码

```
states <- as.data.frame(state.x77[, c("Murder", "Population",
                              "Illiteracy", "Income", "Frost")])
murder_fit <- lm(Murder ~ Population + Illiteracy + Income +
                                           Frost, data = states)

plot(x=fitted(murder_fit),y=rstudent(murder_fit),ylim=c(-4,4))
abline(h=3,col="red",lty=2)
abline(h=-3,col="red",lty=2)

abline(h=2,col="blue",lty=2)
abline(h=-2,col="blue",lty=2)

which(abs(rstudent(murder_fit))>3)

library(car)
outlierTest(murder_fit)        # car 包中的异常点检验函数
```

```
# 高杠杆指点
# 超过2倍或3倍的平均杠杆值即可认为是高杠杆点，这里把Alaska和California作为高杠杆点。
hat.plot <- function(fit){
  p <- length(coefficients(fit))
  n <- length(fitted(fit))
  plot(hatvalues(fit),main = "Index Plot of Hat Values")
  abline(h=c(2,3)*p/n,col="red",lty=2)
  identify(1:n, hatvalues(fit), names(hatvalues(fit)))  #这句产生交互效果，选中某个点后，关闭后返回点的名称
}
hat.plot(murder_fit)

# 强影响点 强影响点是那种若删除则模型的系数会产生明显的变化的点。
#          一种方法是计算Cook距离，一般来说， Cook’s D值大于4/(n-k -1),
#          则表明它是强影响点，其中n 为样本量大小， k 是预测变量数目。
cutoff <- 4/(nrow(states-length(murder_fit$coefficients)-2)) #coefficients加上了截距项，因此要多减1
plot(murder_fit,which=4,cook.levels = cutoff)
abline(h=cutoff,lty=2,col="red")
```