第四章：违背基本假设的情况

马学俊(主讲) 杜悦(助教)

苏州大学

数学科学学院

https://xuejunma.github.io/

Outline

Gauss-Markov 条件

$$\begin{cases} E(\epsilon_i) = 0, \quad i = 1, 2, \ldots, n \\ cov(\epsilon_i, \epsilon_j) = \begin{cases} \sigma^2, i = j \\ 0, i \neq j \end{cases} \quad i, j = 1, 2, \ldots, n \end{cases}$$

- 异方差:

$$var(\varepsilon_i) \neq var(\varepsilon_j), \quad i \neq j$$

- 自相关性：

$$cov(\varepsilon_i, \varepsilon_j) \neq 0 \;\; i \neq j$$

异方差性产生的原因

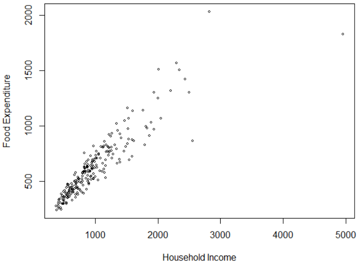
- 某一因素或某些因素随着自变量观察值得变化而对因变量产生不同的影响。

例 (4.1)

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \;\; i = 1, 2, \ldots, n$$

- y_i: 消费、x_i: 收入
- 由于各户的收入、消费观念和习惯不同
- 低收入家庭购买的差异比较小，大多数购买生活必需品
- 高收入家庭的购买邢炜差异比较大：房子、汽车和股票等

低收入的家庭购买差异性比较小,高收入的家庭购买行为差异就很大。导致消费模型的随机项ϵ_i 具有不同的方差。



R

```
> rm(list=ls())
> library(quantreg)
> data(engel)
> head(engel)
    income  foodexp
1 420.1577 255.8394
2 541.4117 310.9587
> attach(engel)
> plot(income,foodexp,xlab="Household Income",
       ylab="Food Expenditure", cex=.5)
```

异方差性带来的问题

当存在异方差时，普通最小二乘估计存在以下问题：

(1) 参数估计值虽是无偏的，但不是最小方差线性无偏估计；

(2) 参数的显著性检验失效；

(3) 回归方程的应用效果极不理想。

模拟R代码

```
> rm(list=ls())
> library(MASS)
> p <- 3; n <- 50
> mu <- rep(0, p)
> sigmu <- diag(rep(1, p))
> x <- MASS::mvrnorm(n=n, mu=mu, Sigma = sigmu)
> sigma <- rep(1:10, length=n)
> e <- rnorm(n=n, mean=0, sd=sigma)
> beta0 <- c(1.5, 2, 1.5)
> y <-  x %*% beta0 + x[, 1]* e
> plot(x[, 1], y)
> fit <- lm(y~x-1)
> summary(fit)
Coefficients:
   Estimate Std. Error t value Pr(>|t|)
x1  -0.2041     0.8627  -0.237   0.8140
x2   1.8931     1.0092   1.876   0.0669 .
x3   1.0701     0.9453   1.132   0.2634
Residual standard error: 6.405 on 47 degrees of freedom
Multiple R-squared:  0.08127, Adjusted R-squared:  0.02262
F-statistic: 1.386 on 3 and 47 DF,  p-value: 0.2587
```

异方差性的检验

- 残差图分析法

Figure: 无异方差

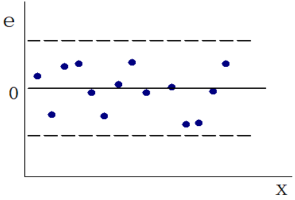
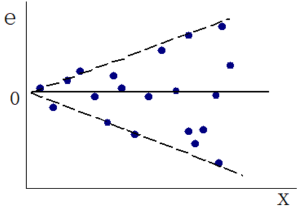


Figure: 存在异方差



异方差性的检验

- 等级相关系数法

等级相关系数检验法又称斯皮尔曼(Spearman) 检验，是一种应用较广泛的方法。这种检验方法既可用于大样本，也可用于小样本。进行等级相关系数检验通常有三个步骤。

第(1)步 作y 关于x 的普通最小二乘回归，求出ϵ_i 的估计值，即e_i 的值。

第(2)步 取e_i 的绝对值，分别把x_i 和$|e_i|$ 按递增（或递减）的次序分成等级，按下式计算出等级相关系数

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

式中n 为样本量；d_i 为对应于x_i 和$|e_i|$ 的等级差数。

异方差性的检验

- 等级相关系数法

第(3)步 做等级相关系数的显著性检验。在$n > 8$ 的情况下,用下式对样本等级相关系数r_s 进行t 检验。检验统计量为:

$$t = \frac{\sqrt{n-2}r_s}{\sqrt{1-r_s^2}}$$

- 如果$|t| \leq t_{\alpha/2}(n-2)$ 可认为异方差性问题不存在;
- 如果$|t| > t_{\alpha/2}(n-2)$,说明x_i 和$|e_i|$ 之间存在系统关系,异方差性问题存在。

例4.3 R代码

```
> rm(list=ls())
> ex43 <- read.table("ex43.txt", head=TRUE, fileEncoding="utf8"
> attach(ex43)
> head(ex43)
      y     x
1 5081 25669
2 2724 17885
> fit43 <- lm(y~x)
> summary(fit43)

             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.274e+02  2.744e+02   0.464    0.646
x           1.068e-01  8.573e-03  12.454 3.66e-13 ***

> cor.test(x=x, y =abs(fit43$residuals), method = "spearman")
S = 2100, p-value = 0.0008485
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.5766129
```

模拟R代码

```r
rm(list=ls())
library(MASS)
p <- 3; n <- 50
mu <- rep(0, p)
sigmu <- diag(rep(1, p))
x <- MASS::mvrnorm(n=n, mu=mu, Sigma = sigmu)
sigma <- rep(1:10, length=n)
e <- rnorm(n=n, mean=0, sd=sigma)
beta0 <- c(1.5, 2, 1.5)
y <-  x %*% beta0 + x[, 1]* e
plot(x[, 1], y)
fit <- lm(y~x-1)
summary(fit)
plot(fit$residuals)
cor.test(x=x[, 1], y =abs(fit$residuals), method = "spearman")
cor.test(x=x[, 2], y =abs(fit$residuals), method = "spearman")
cor.test(x=x[, 3], y =abs(fit$residuals), method = "spearman")
```

一元加权最小二乘估计

消除异方差性的方法通常有：

- 加权最小二乘法, Box-Cox 变换法,(参考文献[1])
- 方差稳定性变换法

加权最小二乘法(Weighted Least Square, 简记为WLS) 是一种最常用的消除异方差性的方法。

一元加权最小二乘估计

一元线性回归普通最小二乘法的残差平方和为：

$$Q(\beta_0,\beta_1)=\sum_{i=1}^{n}(y_i-E(y_i))^2=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2$$

一元线性回归的加权最小二乘的离差平方和为：

$$Q(\beta_0,\beta_1)=\sum_{i=1}^{n}\omega_i(y_i-E(y_i))^2=\sum_{i=1}^{n}\omega_i(y_i-\beta_0-\beta_1x_i)^2$$

加权最小二乘估计为

$$\begin{cases}\hat{\beta}_{0\omega}=\bar{y}_\omega-\hat{\beta}_{1\omega}\bar{x}_\omega\\ \hat{\beta}_{1\omega}=\dfrac{\sum_{i=1}^{n}\omega_i(x_i-\bar{x}_\omega)(y_i-\bar{y}_\omega)}{\sum_{i=1}^{n}\omega_i(x_i-\bar{x}_\omega)^2}\end{cases}$$

其中$\bar{x}_\omega=\frac{1}{\sum\omega_i}\omega_ix_i$ 为自变量的加权平均，$\bar{y}_\omega=\frac{1}{\sum\omega_i}\omega_iy_i$ 为自变量的加权平均。

一元加权最小二乘估计

为了消除异方差的影响，观测值的权数应该是观测值误差项方差的倒数,即

$$\omega_i = \frac{1}{\sigma_i^2}$$

σ_i^2 为第i个观测值误差项方差。误差项方差较大的观测值接受较小的权数；误差项方差较小的观测值接受较大的方差。

在社会经济研究中，经常会遇到误差项方差与x 的幂函数x^m 成比例，其中，m为待定未知参数

$$\omega_i = \frac{1}{x_i^m}$$

一元加权最小二乘估计注意事项

- 加权最小二乘是以牺牲大方差项的拟合效果为代价改善了小方差项的拟合效果，这也并不总是研究者所需要的。
- 在社会经济现象中，通常变量取值大时方差也大，在以经济总量为研究目标时，更关心的是变量取值大的项，而普通最小二乘恰好能满足这个要求
- 在时间序列数据建模时，近期的经济数据往往数值偏大，早期的数据数值会偏小，表现出异方差。
- 对这样的数据使用加权最小二乘，会对早期的数据拟合的更好，而近期的数据拟合效果变差。
- 在一些特定场合下，即使数据存在异方差，也仍然可以选择使用普通最小二乘估计。

多元加权最小二乘法

当误差项ϵ_i 存在异方差性时，对于一般的多元线性回归模型，加权离差平方和为

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

记

$$\boldsymbol{W} = \begin{pmatrix} \omega_1 & & & \vdots \\ & \omega_2 & & \\ & & \ddots & \\ \vdots & & & \omega_n \end{pmatrix}$$

加权最小二乘估计的矩阵表达式为

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

多元加权最小二乘法：权函数的确定方法

通常取权函数$\boldsymbol{W}$ 为某个自变量$x_j, j = 1, 2, \ldots, p)$ 的幂函数，即$\boldsymbol{W} = x_j^m$

★ 在$x_1, x_2, \cdots, x_p$ 这p 个自变量中取哪一个？

这只需计算每个自变量x_j 与普通残差的等级相关系数，选取等级相关系数最大的自变量构造权函数。例4.4

自相关性产生的背景和原因

如果一个回归模型的随机误差项$cov(\epsilon_i, \epsilon_j) \neq 0$则称随机误差项之间存在着自相关现象。

这里的自相关现象不是指两个或两个以上的变量之间的相关,而指的是一个变量前后期数值之间存在的相关关系。

自相关性产生的背景和原因

(1) 遗漏关键变量时会产生序列的自相关性。
(2) 经济变量的滞后性会给序列带来自相关性。
(3) 采用错误的回归函数形式也可能引起自相关性。
(4) 蛛网现象(Cobweb phenomenon) 可能带来序列的自相关性。
(5) 因对数据加工整理而导致误差项之间产生自相关性。

自相关性带来的问题

(1) 参数的估计值不再具有最小方差线性无偏性。

(2) 均方误差MSE 可能严重低估误差项的方差。

(3) 容易导致对t 值评价过高,常用的F 检验和t 检验失效。如果忽视这一点,可能导致得出回归参数统计检验为显著,但实际上并不显著的严重错误结论。

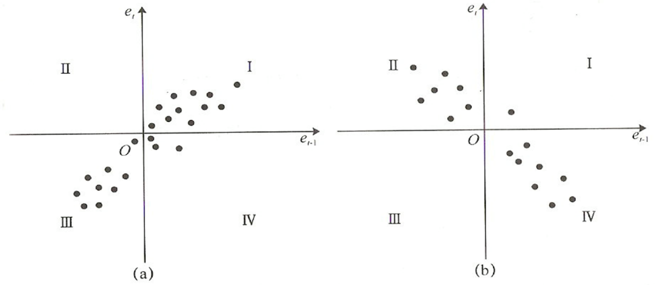
(4) 当存在序列相关时,仍然是 β 的无偏估计量,但在任一特定的样本中,可能严重歪曲 β 的真实情况,即最小二乘估计量对抽样波动变得非常敏感

(5) 如果不加处理地运用普通最小二乘法估计模型参数,用此模型进行预测和结构分析将会带来较大的方差甚至错误的解释。

自相关性的诊断

- 图示检验法

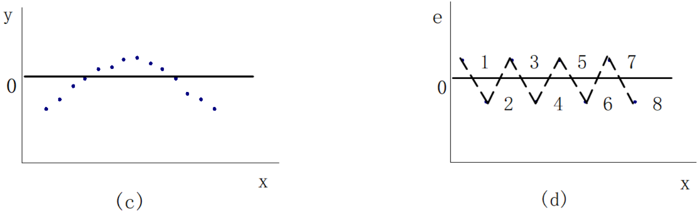
(1) 绘制(e_t, e_{t-1})的散点图。



自相关性的诊断

- 图示检验法

(2) 按照时间顺序绘制回归残差项e_t 的图形



自相关性的诊断

- 自相关系数法

误差序列$\epsilon_1, \epsilon_2, \cdots, \epsilon_n$ 的自相关系数定义为:

$$\rho = \frac{\sum_{t=2}^{n} \epsilon_t \epsilon_{t-1}}{\sqrt{\sum_{t=2}^{n} \epsilon_t^2}\sqrt{\sum_{t=2}^{n} \epsilon_{t-1}^2}}$$

ρ 的取值范围是$[-1, 1]$，当ρ 接近1 时，表明序列误差存在正相关，当ρ 接近-1 时，表示序列误差存在负相关。
在实际应用中，误差序列$\epsilon_1, \epsilon_2, \cdots, \epsilon_n$ 的值是未知的，需要用其估计值e_i 代替，得到自相关系数的估计值为

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

自相关性的诊断

- DW 检验

DW检验是J.Durbin 和G.S.Watson 于1951 年提出的一种适用于小样本的一种检验方法。
DW检验只能用于检验随机扰动项具有一阶自回归形式的序列相关问题。这种检验方法是建立计量经济学模型中最常用的方法,一般的计算机软件都可自动产生出D.W 值。
随机误差项的一阶自回归形式为

$$\epsilon_t = \rho\epsilon_{t-1} + \mu_t$$

为了检验序列的相关性，构造的假设是

$$H_0: \quad \rho = 0$$

自相关性的诊断

- DW 检验

定义DW统计量为：

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

如果认为$\sum_{t=2}^{n} e_t^2$ 与$\sum_{t=2}^{n} e_{t-1}^2$ 近似相等

$$DW = \frac{\sum_{t=2}^{n} e_t^2 + \sum_{t=2}^{n} e_{t-1}^2 - 2\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e_t^2} \approx 2\left[1 - \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e_t^2}\right]$$

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

因此

$$DW \approx 2(1 - \hat{\rho})$$

DW 检验

Figure: DW 值与$\hat{\rho}$ 对应关系

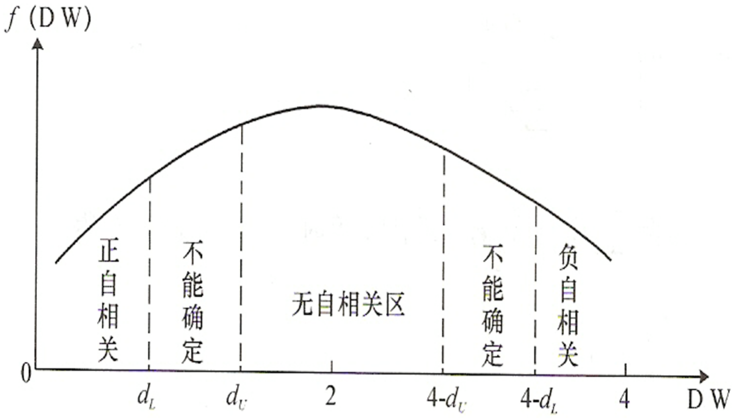
$\hat{\rho}$	D.W	误差项的自相关性
−1	4	完全负自相关
(−1，0)	(2，4)	负自相关
0	2	无自相关
(0，1)	(0，2)	正自相关
1	0	完全正自相关

DW 检验

根据样本容量n和解释变量的数目k(这里包括常数项),查DW 分布表,得临界值d_L 和d_U, 然后依下列准则考察计算得到的DW值,以决定模型的自相关状态:

$0 \leqslant D.W \leqslant d_L$,	误差项 $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$间存在正相关;
$d_L < D.W \leqslant d_U$,	不能判定是否有自相关;
$d_U < D.W < 4-d_U$,	误差项 $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$间无自相关;
$4-d_U \leqslant D.W < 4-d_L$,	不能判定是否有自相关;
$4-d_L \leqslant D.W \leqslant 4$,	误差项 $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$间存在负相关。

DW 检验



DW 检验

D.W检验尽管有着广泛的应用,但也有明显的缺点和局限性。

- DW 检验有一个不能确定的区域,一旦DW 值落在这个区域,就无法判断。这时,只有增大样本容量或选取其他方法。
- DW 统计量的上、下界表要求$n > 15$, 这是因为样本如果再小,利用残差就很难对自相关的存在性作出比较正确的诊断。
- DW 检验不适应随机项具有高阶序列相关的检验。

自相关问题的处理方法

- 迭代法

以一元线性回归模型为例，设一元线性回归模型的误差项存在一阶自相关

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t$$

$$\epsilon_t = \rho\epsilon_{t-1} + \mu_t$$

$$\begin{cases} E(\mu_t) = 0, \quad t = 1, 2, \ldots, n \\ cov(\mu_t, \mu_s) = \begin{cases} \sigma^2, t = s \\ 0, t \neq s \end{cases} \quad t, s = 1, 2, \ldots, n \end{cases}$$

根据一元线性回归模型$y_t = \beta_0 + \beta_1 x_t + \epsilon_t$，有

$$y_{t-1} = \beta_0 + \beta_1 x_{t-1} + \epsilon_{t-1}$$

迭代法

变形后有

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

令：

$$y'_t = y_t - \rho y_{t-1}$$

$$x'_t = x_t - \rho x_{t-1}$$

$$\beta'_0 = \beta_0(1 - \rho)$$

$$\beta'_1 = \beta_1$$

得到有随机独立误差项，满足线性回归基本假设的

$$y'_t = \beta'_0 + \beta'_1 x'_t + \mu_t \qquad (*)$$

迭代法

其中自相关系数ρ 用公式$\hat{\rho} \approx 1 - \frac{1}{2}DW$ 估计。

用变换因变量与变换自变量作普通最小二乘回归。如果误差项确实是一阶自相关，通过以上变换，回归模型已经消除自相关。

实际问题中，有时误差项并不是简单的一阶自相关，而是更复杂的自相关形式，$(*)$ 式的误差项u_t 可能仍然存在自相关，这就需要进一步对$(*)$ 式的误差项u_t 做DW 检验，以判断u_t 是否存在自相关，如果检验表明误差项u_t 不存在自相关，迭代法到此结束。如果检验表明误差项u_t 存在自相关，那末对回归模型$(*)$ 式重复用迭代法，这个过程可能要重复几次，直至最终消除误差项自相关。这种迭代消除自相关的过程正是迭代法名称的由来。

自相关问题的处理方法

- 差分法

一阶差分法通常适用于原模型存在较高程度的一阶自相关的情况。在迭代法中，当ρ

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

为:

$$(y_t - \rho y_{t-1}) = \beta_1(x_t - x_{t-1}) + (\epsilon_t - \epsilon_{t-1})$$

以$\Delta y_t = y_t - y_{t-1}$，$\Delta x_t = x_t - x_{t-1}$，得到

$$\Delta y_t = \beta_1 \Delta x_t + \mu_t$$

上式是不带有常数项的回归方程

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

差分法

一阶差分法的应用条件是自相关系数$\rho = 1$，在实际应用中，ρ 接近1 时我们就采用差分法而不用迭代法，这有两个原因。

第一，迭代法需要用样本估计自相关系数ρ，对ρ 的估计误差会影响迭代法的使用效率；

第二，差分法比迭代法简单,人们在建立时序数据的回归模型时,更习惯于用差分法。

但是完全的$\rho = 1$ 情况几乎是见不到的,实际应用时ρ 较大就行!

BOX-COX 变换

BOX-COX 变换是由博克斯（BOX）与考克斯（COX）在1964 年提出的一种应用非常广泛的变换，它是对因变量y做如下变换：

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

其中，λ 为待定参数。此变换要求y 的各分量都大于0。否则可用下面推广的BOX-COX 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 做平移，使得$y + a$ 的各个分量都大于0 后再做BOX-COX 变换。

对于不同的λ，所做的变换也不同，所以这是一个变换族。它包含一些常用的变换，如对数变换$(\lambda = 0)$，平方根变换$(\lambda = 1/2)$ 和倒数变换$(\lambda = -1)$。

BOX-COX 变换

寻找合适的λ，使得变换后

$$\boldsymbol{y}^{(\lambda)} = \begin{pmatrix} y_1^{(\lambda)} \\ y_2^{(\lambda)} \\ \vdots \\ y_n^{(\lambda)} \end{pmatrix} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$$

从而符合线性回归模型的各项假设：误差分量等方差、不相关等。

事实上，BOX-COX 变换不仅可以处理异方差性问题，还能处理自相关、误差非正态、回归函数非线性等情况。

BOX-COX 变换

经过计算可得λ 的最大似然估计（参见参考文献[2]）

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

式中，$\hat{\sigma}_\lambda^2 = \frac{1}{n} SSE(\lambda, y^{(\lambda)})$，$|\boldsymbol{J}| = \prod_{i=1}^n |\frac{dy_I^{(\lambda)}}{dy_i}| = \prod_{i=1}^n y_i^{(\lambda-1)}$

令$z^{(\lambda)} = \frac{y^{(\lambda)}}{|\boldsymbol{J}|}$，对$L_{max}$ 取对数并略去与λ 无关的常数项，可得

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

为找出λ，使得$\ln L_{max}(\lambda)$ 达到最大，只需使$SSE(\lambda, z^{(\lambda)})$ 达到最小即可。它的解析解比较难找，通常是给出一系列λ 的值，计算对应的$SSE(\lambda, z^{(\lambda)})$，取使得$SSE(\lambda, z^{(\lambda)})$ 达到最小的λ 即可。

BOX-COX 变换

- BOX-COX 变换是一个幂变换族，其中当变换参数 $\lambda=0$ 时成为对数变换，而对数变换则是比幂变换应用更广泛的变换，很多场合都可以首先尝试对数据作对数变换。
- 从概率分布的角度看，当数据本身服从对数正态分布时，对数据取对数变换后就服从正态分布。对数正态分布是右偏分布，有厚重的右尾。
- 从数据看，如果数据中一些数值很大，但是小数值的数据更密集，个数也更多，大数值的数据较较疏松，个数较少，这样的数据很可能服从对数正态分布，可以尝试对数变换。
- 对回归分析问题，如果只对因变量作对数变换，就是BOX-COX 变换 $\lambda=0$时的特例。也可以考虑只对自变量作对数变换，或者同时对因变量和对自变量作对数变换。

异常值与强影响点

异常值分为两种情况：

- 一种是关于因变量y 异常；
- 另一种是关于自变量x 异常。

关于因变量y的异常值

在残差分析中，认为超过$\pm 3\hat{\sigma}$ 的残差为异常值。标准化残差

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

学生化残差

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

h_{ii} 为$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ 的主对角线元素。

当观测数据中存在关于y的异常观测值时，普通残差、标准化残差、学生化残差这三种残差都不再适用。这是由于异常值把回归线拉向自身，使异常值本身的残差减小，而其余观测值的残差增大，这时回归标准差$\hat{\sigma}$ 也会增大，因而用传统的“3σ”准则不能正确分辨出异常值。解决这个问题的方法是改用删除残差。

关于因变量y的异常值

删除残差的构造思想是：在计算第i个观测值的残差时，用删除掉的第i个观测值的其余$n-1$ 个观测值拟合回归方程，计算出第i个观测值的删除拟合值$\hat{y}_{(i)}$，这个删除拟合值与第i 个值无关，不受第i 个值是否为异常值的影响，由此定义第i 个观测值的删除残差为

$$e_{(i)} = y_i - \hat{y}_{(i)}$$

可以证明

$$e_{(i)} = \frac{e_i}{1-h_{ii}}$$

进一步，可以给出第i 个观测值的删除学生残差，记为$ser_{(i)}$。

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

$|SRE_{(i)}| > 3$ 的观测值即判定为异常值。

关于自变量x的异常值对回归的影响

在$D(e_i)=(1-h_{ii})\sigma^2$ 中，h_{ii} 为帽子矩阵中主对角线的第i个元素，它是调节e_i 方差大小的杠杆，因而称h_{ii} 为第i个观测值的杠杆值。类似于一元线性回归，多元线性回归的杠杆值h_{ii} 也表示自变量的第i次观测与自变量平均值之间距离的远近。较大的杠杆值的残差偏小，这是因为杠杆值大的观测点远离样本中心，能够把回归拉向自身，因而把杠杆值大的样本点称为强影响点。

因为$tr(\boldsymbol{H})=\sum_{i=1}^{n}h_{ii}=p+1$，则杠杆值的平均值为

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

一个杠杆值h_{ii} 大于2 倍或者3倍的$\bar{h}$，就认为是大的。

计算中心化杠杆值ch_{ii}，由文献[2] 可知

$$ch_{ii}=h_{ii}-1/n$$

因此$\sum_{i=1}^{n}ch_{ii}=p$，中心化杠杆值得平均值是

$$\bar{ch}=\frac{1}{n}\sum_{i=1}^{n}ch_{ii}=\frac{p}{n}$$

关于自变量x的异常值对回归的影响

- 虽然强影响点并不总是y的异常值点，不能单纯根据杠杆值h_{ii} 的大小判断强影响点是否异常，但是我们对强影响点应该有足够的重视。
- 为此引入库克距离，用来判断强影响点是否为y 的异常值点。库克距离的计算公式为：

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

- 库克距离反映了杠杆值h_{ii} 与残差e_i 的综合效应。对于库克距离，判断其大小的方法比较复杂，一个粗略的标准是:
  - 当$D_i < 0.5$ 时，认为不是异常值点，
  - 当$D_i > 1$ 时，认为是异常值点。

异常值产生的原因和消除方法

异常值原因	异常值消除方法
1. 数据登记误差，存在抄写或录入的错误	重新核实数据
2. 数据测量误差	重新测量数据
3. 数据随机误差	删除或重新观测异常值数据
4. 缺少重要自变量	增加必要的自变量
5. 缺少观测数据	增加观测数据，适当扩大自变量取值范围
6. 存在异方差	采用加权线性回归
7. 模型选用错误，线性模型不适用	改用非线性回归模型

作业

- p.124 4.4
- p.124 4.9
- p.124 4.14