回归分析:最小二乘渐进性质

学业辅导中心

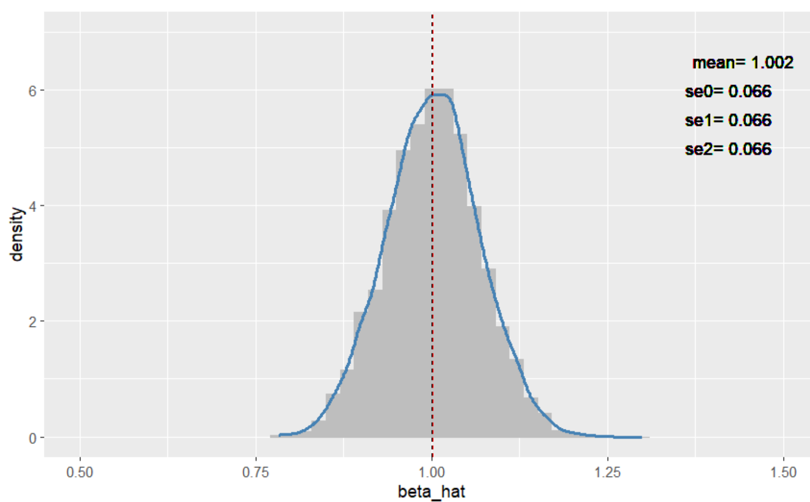
一般的统计软件包会告诉我们, 在normal linear model中统计量的点估计, 标准误和p值对于某一个β_j. 即

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

然而这个假定非常强, 它要求我们

- 有线性的形式
- 误差项与X独立
- 所有误差项独立同分布于一个正态分布(同方差)

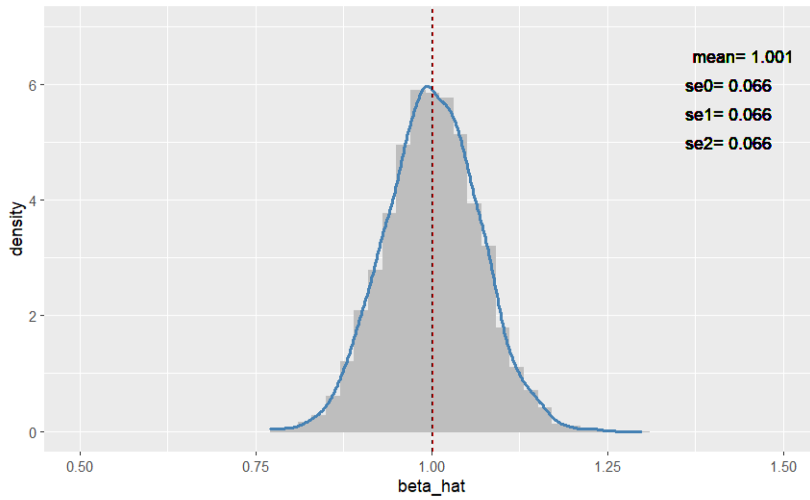
我们做一些模拟试验来放宽这些假设.
我们生成5000次数据, 第一次我们生成用理想的正态分布.

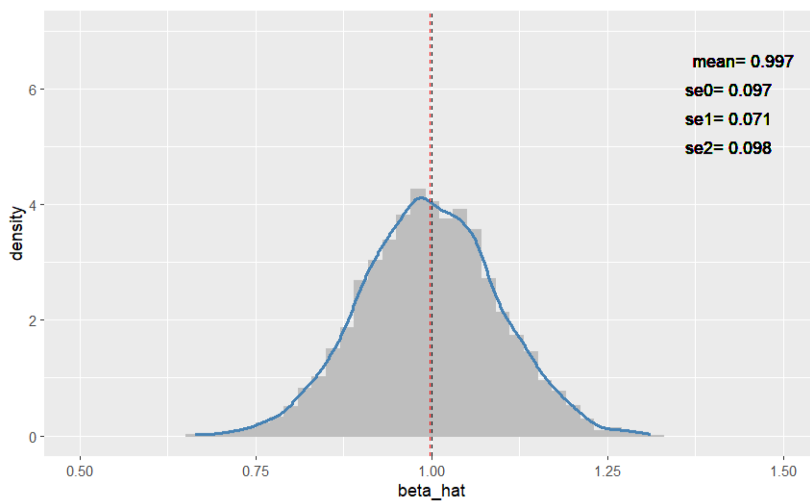


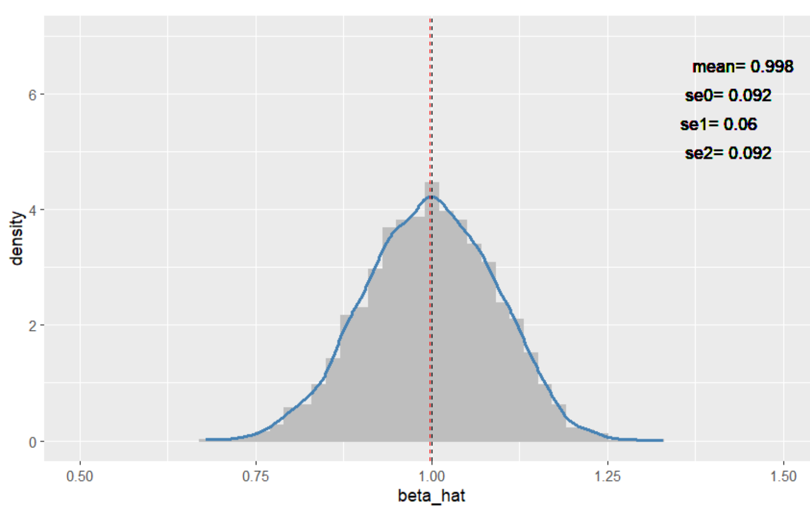
现在我们假设误差是指数分布再做一次OLS
第三幅图是设定不同方差(与X有关)但是是正态分布
第四幅图是设定不同方差(与X有关)但是是均匀分布
我们把这四张图合在一起

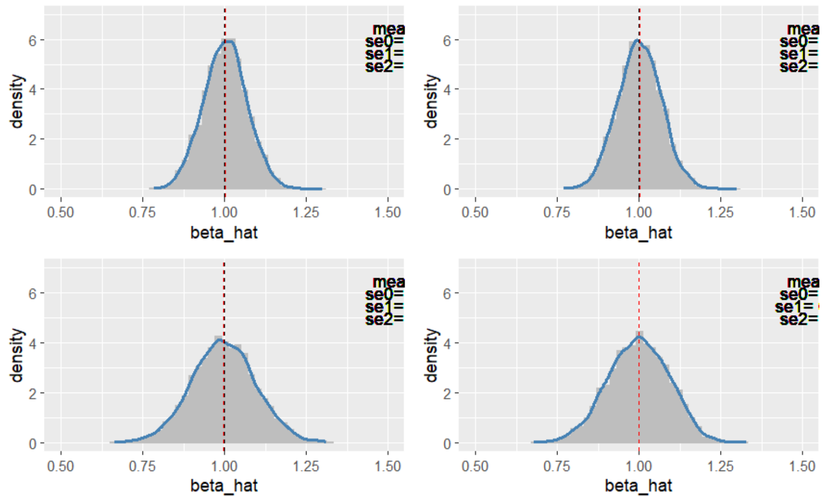
注记. *这四张图中: 都是se2和se0接近, se2是异方差假设下的方差估计, 因此不同于同方差假设下的方差估计se1.*

这提示我们, 在normal linear model中, 很难处理异方差问题.


mean= 1.001
se0= 0.066
se1= 0.066
se2= 0.066
density
beta_hat


mean= 0.997
se0= 0.097
se1= 0.071
se2= 0.098
density
beta_hat


mean= 0.998
se0= 0.092
se1= 0.06
se2= 0.092
density
beta_hat



1 异方差线性模型

定义 (异方差(heteroskedasitic)线性模型)**.**

$$y_i = x_i^{\mathrm{T}}\beta + \varepsilon_i,$$

其中, ε_i相互独立, 且均值为0, 方差为σ_i^2. 设计矩阵X是固定的, 且是列满秩的. $(\beta, \sigma_1^2, \cdots, \sigma_n^2)$是未知参数.

注记.
- 误差项有不同的方差, 且可以依赖于x_i;
- 不做额外的正态性假设, 因此我们不研究OLS估计都分布;
- 需要用到渐进分析理论.

1.1 基本渐进理论

定义 (依概率收敛(coverge to ... in probability))**.** $Z_n, Z \in \mathbb{R}^k$.若

$$\mathrm{pr}\{\|Z_n - Z\| > c\} \to 0, \quad n \to \infty \quad (\forall c)$$

则$Z_n \xrightarrow{p} Z$.

注记 (一般收敛也是依概率收敛)**.**

$$Z_n \to Z \Rightarrow Z_n \xrightarrow{p} Z$$

注记 (向量的依概率收敛)**.**

$$Z_n \xrightarrow{p} Z, W_n \xrightarrow{p} W \Rightarrow (Z_n, W_n) \xrightarrow{p} (Z, W)$$

性质 (Khinchin's WLLN)**.** 若$Z_1, \cdots, Z_n \overset{iid}{\sim} F$, 期望为$\mu \in \mathbb{R}^k$, 则$n^{-1}\sum_{i=1}^{n} Z_i \xrightarrow{p} \mu$.

性质 (向量形式的Chebyshev's Inequality)**.** 下列范数为2-范数.

$$P(\|X-EX\|\geq a)\leq\frac{E\|X-EX\|^p}{a^p}.\quad p\geq 1$$

例 1. 若Z_n有零均值向量, 且$cov(Z_n)=a_nC_n, a_n\to 0, C_n\to C<\infty$, 则$Z_n\xrightarrow{P} Z$.

证明.

$$\begin{aligned}\operatorname{pr}\{\|Z_n\|>c\} &\leq c^{-2}E\left\{\|Z_n\|^2\right\}\\ &=c^{-2}E\left(Z_n^{\mathrm{T}}Z_n\right)\\ &=c^{-2}\operatorname{trace}\left\{E\left(Z_nZ_n^{\mathrm{T}}\right)\right\}\\ &=c^{-2}\operatorname{trace}\left\{\operatorname{cov}(Z_n)\right\}\\ &=c^{-2}a_n\operatorname{trace}(C_n)\to 0,\end{aligned}$$

□

定义 (依分布收敛)**.** 若对于分布$F: t\mapsto P(Z\leq t)$全部的连续点z, 满足$P(Z_n\leq z)\to P(Z\leq z), n\to\infty$, 则称$Z_n\xrightarrow{D} Z$.

性质. 在概率论中我们学过, 随机变量依分布收敛于常数等价于依概率收敛于常数. 一般来说, 依分布收敛推不出依概率收敛, 但是依概率收敛能推出依分布收敛.

性质. 若$Z_n\xrightarrow{D} Z, W_n\xrightarrow{P} c$, 则

1. $Z_n+W_n\xrightarrow{D} Z+c$;
2. $Z_nW_n\xrightarrow{D} cZ$;
3. 当$c\neq 0, Z_nW_n^{-1}\xrightarrow{D} c^{-1}Z$

2　最小二乘估计量的相合性(consistency)

性质. 在异方差线性模型下, 最小二乘估计量的表达式是$\hat\beta-\beta=B_n^{-1}\xi_n$, 其中

$$B_n=\frac{1}{n}\sum_{i=1}^n x_ix_i^T,\quad \xi_n=\frac{1}{n}\sum_{i=1}^n x_i\epsilon_i.$$

证明. 在OLS中, $\hat\beta=(X^\top X)^{-1}X^\top Y$, 于是

$$\begin{aligned}\hat\beta &= B_n^{-1}n^{-1}\sum_{i=1}^n x_iy_i\\ &=B_n^{-1}n^{-1}\sum_{i=1}^n x_i(x_i^{\mathrm{T}}\beta+\varepsilon_i)\\ &=B_n^{-1}B_n\beta+B_n^{-1}n^{-1}\sum_{i=1}^n x_i\varepsilon_i\\ &=\beta+B_n^{-1}n^{-1}\sum_{i=1}^n x_i\varepsilon_i.\end{aligned}$$

□

性质. *1.* 最小二乘估计量是无偏的.

证明.
$$E(\hat{\beta}-\beta)=E(B_n^{-1}\xi_n)=B_n^{-1}\sum_{i=1}^{n}x_iE(\varepsilon_i)=0$$

□

2. 最小二乘估计量的协方差矩阵为$\text{cov}(\hat{\beta})=n^{-1}B_n^{-1}M_nB_n^{-1}$. 其中$M_n=n^{-1}\sum_{i=1}^{n}\sigma_i^2x_ix_i^{\mathrm{T}}$.

证明.
$$\text{cov}(\xi_n)=\text{cov}\left(n^{-1}\sum_{i=1}^{n}x_i\varepsilon_i\right)=n^{-2}\sum_{i=1}^{n}\sigma_i^2x_ix_i^{\mathrm{T}}\equiv M_n/n,$$

□

这是这个最小二乘估计量的三明治形式*(Sandwich form)*, B_n是*“bread”*, M_n是*“Meat”*.

2.1 相合性

在这里, 我们引入以下的假设,

假设 1. 有有限的B,M使得$B_n\to B, M_n\to M$, 且B可逆.

性质. 在假设*1*下, 异方差线性模型的*OLS*估计量$\hat{\beta}$是β的相合估计量

证明. 只需证$\xi_n\xrightarrow{P}0$, 事实上,

$$\begin{aligned}P(\|\xi_n\|>c)&\le\frac{1}{c^2}E\|\xi_n\|^2\\&=\frac{1}{c^2}tr(cov(\xi_n))\\&=\frac{1}{nc^2}M_n\to0\quad(n\to\infty)\end{aligned}$$

□

2.2 渐进正态性

渐进正态性需要我们额外加入高阶矩条件.

假设 2.
$$d_{2+\delta,n}=n^{-1}\sum_{i=1}^{n}\|x_i\|^{2+\delta}E(\varepsilon_i^{2+\delta}).$$
存在与n无关的δ和C使得$d_{2+\delta,n}\le C$.

性质. 异方差线性模型, 若满足假设*1*和*2*, 则

$$\sqrt{n}(\hat{\beta}-\beta)\xrightarrow{D}\mathrm{N}(0,B^{-1}MB^{-1})$$

2.3 异方差场合的方差估计: 三明治估计(Sandwich variance estimator)

根据中心极限定理,

$$\hat{\beta} \overset{a}{\sim} \mathrm{N}(\beta, n^{-1}B^{-1}MB^{-1}),$$

- 用ε_i替换σ_i, 于是$\tilde{M}_n = n^{-1}\sum_{i=1}^{n}\varepsilon_i^2 x_i x_i^{\mathrm{T}}$
- 用$\hat{\varepsilon}_i$替换ε_i, 于是$\hat{M}_n = n^{-1}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 x_i x_i^{\mathrm{T}}$.

性质. 当以下几项能被一个常数C限制时,

$$n^{-1}\sum_{i=1}^{n}\mathrm{var}(\epsilon_i^2)x_{ij_1}^2 x_{ij_2}^2, \quad n^{-1}\sum_{i=1}^{n}x_{ij_1}x_{ij_2}x_{ij_3}x_{ij_4}, \quad n^{-1}\sum_{i=1}^{n}\sigma_i^2 x_{ij_1}x_{ij_2}x_{ij_3}$$

在满足假设1时, 异方差线性模型中, $\hat{M}_n \xrightarrow{P} M_n$

定义 (稳健方差估计: EHW方差估计)**.** 元素形式:

$$\hat{V}_{\mathrm{EHW}} = n^{-1}\left(n^{-1}\sum_{i=1}^{n}x_i x_i^{\mathrm{T}}\right)^{-1}\left(n^{-1}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 x_i x_i^{\mathrm{T}}\right)\left(n^{-1}\sum_{i=1}^{n}x_i x_i^{\mathrm{T}}\right)^{-1} = \frac{1}{n}B_n^{-1}\hat{M}_n B_n^{-1}$$

向量形式:

$$\hat{V}_{\mathrm{EHW}} = (X^{\mathrm{T}}X)^{-1}(X^{\mathrm{T}}\hat{\Omega}X)(X^{\mathrm{T}}X)^{-1} \quad \hat{\Omega} = \mathrm{diag}\left\{\hat{\varepsilon}_1^2, \ldots, \hat{\varepsilon}_n^2\right\}$$

注记.

- *Eicher(1967)*第一次使用稳健方差估计
- *White(1980)*将它在经济领域广泛应用
- *Cox(1961)*和*Huber(1867)*考虑了当参数模型中模型误定时的情况
- *Fuller(1975)*提出来在抽样调查中一般形式的稳健方差估计

我们就可以利用渐进正态性做统计推断, 比如说, 我们对于线性假设可以用$\hat{\beta} \overset{a}{\sim} \mathrm{N}(\beta, \hat{V}_{\mathrm{EHW}})$, 对于某些特定的参数可以用$\hat{\beta}_j \overset{a}{\sim} \mathrm{N}(\beta_j, \hat{\mathrm{se}}_{\mathrm{EHW},j}^2)$.

当然, 实际估计的时候, 也有很多修正的情况

$$\hat{V}_{EHW,k} = n^{-1}\left(n^{-1}\sum_{i=1}^{n}x_i x_i^T\right)^{-1}\left(n^{-1}\sum_{i=1}^{n}\hat{\epsilon}_{i,k}^2 x_i x_i^T\right)\left(n^{-1}\sum_{i=1}^{n}x_i x_i^T\right)^{-1}$$

其中,

$$\hat{\epsilon}_{i,k} = \begin{cases} \hat{\epsilon}_i, & \text{(HC0)} \\ \hat{\epsilon}_i\sqrt{\dfrac{n}{n-p}}, & \text{(HC1)} \\ \hat{\epsilon}_i/\sqrt{1-h_{ii}}, & \text{(HC2)} \\ \hat{\epsilon}_i/(1-h_{ii}), & \text{(HC3)} \\ \hat{\epsilon}_i/(1-h_{ii})^{\min\{2, nh_{ii}/(2p)\}}, & \text{(HC4).} \end{cases}$$

2.4 异方差退化为同方差的情况

在同方差场合, $\sigma_i^2=\sigma^2$. 我们有$M_n=\sigma^2 n^{-1}\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}=\sigma^2 B_n$ 因此

$$\operatorname{cov}(\hat{\beta})=\frac{1}{n}B_n^{-1}M_nB_n^{-1}=n^{-1}\sigma^2B_n^{-1}$$

以及$\sqrt{n}(\hat{\beta}-\beta)\xrightarrow{D}\mathrm{N}(0,\sigma^2B^{-1})$.

$$\hat{V}=\hat{\sigma}^2\left(\sum_{i=1}^{n}x_ix_i^{\mathrm{T}}\right)^{-1}=\hat{\sigma}^2\left(X^{\mathrm{T}}X\right)^{-1}$$
$$\hat{\beta}\overset{\mathrm{a}}{\sim}\mathrm{N}\left(\beta,\hat{\sigma}^2\left(X^{\mathrm{T}}X\right)^{-1}\right)$$

且$\hat{\sigma}^2$是σ^2的相合估计量(因为$\hat{M}_n$是M的相合估计). $V\hat{a}r_{EHW}=\frac{1}{n}B_n^{-1}M_nB_n^{-1}=\frac{1}{n}B_n^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2x_ix_i^{\mathrm{T}}\right)B_n^{-1}=\frac{\hat{\sigma}^2}{n}B_n^{-1}$.

3 应用

3.1 齐方差

```
> round(lalonde.t, 3)
               OLS    hc0    hc1    hc2    hc3    hc4
(Intercept)  0.073  0.070  0.069  0.069  0.067  0.066
age          1.170  1.294  1.276  1.271  1.248  1.249
educ         1.751  2.032  2.005  1.988  1.943  1.915
black       -1.736 -1.999 -1.972 -1.953 -1.907 -1.905
hisp         0.272  0.304  0.300  0.296  0.289  0.288
married     -0.166 -0.171 -0.169 -0.168 -0.164 -0.163
nodegr      -0.015 -0.015 -0.014 -0.014 -0.014 -0.014
re74         1.405  0.976  0.963  0.920  0.866  0.773
re75         0.131  0.139  0.137  0.134  0.129  0.122
u74          1.162  0.890  0.878  0.868  0.847  0.832
u75         -1.045 -0.761 -0.751 -0.749 -0.737 -0.743
treat        2.606  2.490  2.456  2.449  2.407  2.404
```

3.2 异方差

```
> summary(boston.lm)

Call:
lm(formula = medv ~ ., data = BostonHousing)

Residuals:
    Min      1Q  Median      3Q     Max
-15.595  -2.730  -0.518   1.777  26.199

Coefficients:
```

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
zn           4.642e-02  1.373e-02   3.382 0.000778 ***
indus        2.056e-02  6.150e-02   0.334 0.738288
chas1        2.687e+00  8.616e-01   3.118 0.001925 **
nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
age          6.922e-04  1.321e-02   0.052 0.958229
dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
b            9.312e-03  2.686e-03   3.467 0.000573 ***
lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0  ‘’ *** 0.001  ‘’ ** 0.01  ‘’ * 0.05  ‘’ . 0.1  ‘ ’  1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

```
> round(boston.t, 3)
                OLS    hc0    hc1    hc2    hc3    hc4
(Intercept)   7.144  4.621  4.557  4.477  4.334  4.247
crim         -3.287 -3.784 -3.732 -3.478 -3.166 -2.584
zn            3.382  3.420  3.372  3.345  3.271  3.276
indus         0.334  0.414  0.408  0.406  0.398  0.401
chas1         3.118  2.106  2.077  2.051  1.997  1.997
nox          -4.651 -4.759 -4.693 -4.643 -4.528 -4.516
rm            9.116  4.573  4.509  4.426  4.281  4.184
age           0.052  0.043  0.042  0.042  0.040  0.040
dis          -7.398 -6.969 -6.872 -6.812 -6.657 -6.657
rad           4.613  5.052  4.982  4.908  4.762  4.653
tax          -3.280 -4.649 -4.584 -4.540 -4.432 -4.415
ptratio      -7.283 -8.227 -8.113 -8.060 -7.894 -7.927
b             3.467  3.525  3.476  3.435  3.344  3.296
lstat       -10.347 -5.340 -5.266 -5.176 -5.014 -4.932
```