第四章 异常情况下的多元回归分析

- 4．1．虚拟变量进行回归分析
- 4．2．异方差(Heteroscedasticity)
- 4．3．模型的多重共线性问题
- 4．4．有关模型设定和数据的问题
- 4．5．数据缺失，非随即抽样和异常值的处理

1

4．1．虚拟变量进行回归分析

- 4．1．1 对定性信息的描述
- 定性因素通常采用二元取值的形式
- 虚拟变量(dummy variables)
- 4．1．2只有一个自变量是虚拟变量

$$threeceo = \beta_0 + \delta_0 indc + \beta_1 asset + \beta_2 eachearn + \varepsilon$$

- 属于制造业时取indc=1，其他行业时=0
- 相同公司规模和盈利的公司CEO年薪相同

$$\delta_0 = E(wage \mid indc = 1, asset, earn, hold) - E(wage \mid indc = 0, asset, earn, hold)$$

- 制造业平均年薪是截距项，其他行业是 $\beta_0 + \delta_0$

2

- 包含多个自变量的情形，只要加入其它变量

$$threeceo = \beta_0 + \delta_0 indc + \beta_1 asset + \beta_2 eachearn + \beta_3 first + \beta_4 second + \varepsilon$$

- 行业对年薪没有影响的零假设为：$H_0 : \delta_0 = 0$

$$threeceo = 78.49 - 10.60 indc + 4.94(E-6) asset + 46.91 learn - 12.74 first + 80.10 second$$

（8.234）（5.273）（1.12E-6）（5.054）（18.20）（36.24）

- 如果什么也不控制我们有

$$\ln threeceo = 97.897 - 13.761 indc$$

（4.255）（5.669）

- 由于没有控制其它因素，给出的年薪差异也比较大，所以前一个其它条件相同的结果更可信

3

- 虚拟变量作为自变量通常用在由于某一特性而带来的因果关系分析或策略分析
- 因变量取对数形式，自变量中出现虚拟变量
- 系数表示的是比率,当系数比较大时还需要考虑二次项的影响

$$\ln threeceo = 8.71 - 0.177 indc + 0.220 \ln asset + 0.354 eachearn$$

（0.325）（0.047）（0.015）（0.045）

$$- 0.357 first + 0.628 second$$

（0.162）（0.320）

- 控制了上述因素之后制造业比其他行业的年薪要低17.7%，采用百分比为16.3%

4

- 4．1．3 多个类别使用虚拟变量

$$\ln threeceo = 8.922 - 0.155 indc + 0.265 indk + 0.046 indm$$

（0.344）（0.050）（0.131）（0.097）

$$+ 0.186 \ln asset + 0.396 eachearn + 0.031 \ln income - 0.404 first$$

（0.017）（0.046）（0.010）（0.161）

$$+ 0.581 second - 0.171 dirchange - 0.039 yreturn$$

（0.316）（0.057）（0.017）

- 所有的系数都在5%置信水平的t检验下显著，基准组为其他行业
- 一个回归模型中我们可以同时使用几个虚拟变量
- 回归模型中包含g个组类，我们需要在模型中包含g-1个虚拟变量

5

- 4．1．4 使用虚拟变量来设置顺序信息
- 假定地方政府的债券评定的等级为从0到4，0代表信用等级较低，4代表最好的信用等级
- 最简单的方式是把CR与其它自变量一样看待

$$MBR = \beta_0 + \beta_1 CR + 其它因素$$

- β_1表示CR增加一个等级对MBR改变的百分比
- 只能说CR=4优于CR=3，它们间的差别到底是什么？
- 比较好的方式是对每一个等级设定一个虚拟变量
- 记 CR_1=1，当CR=1；CR_1 =0，当CR为其它值

$$MBR = \beta_0 + \delta_1 CR_1 + \delta_2 CR_2 + \delta_3 CR_3 + \delta_4 CR_4 + otherfactor$$

6

- δ_1为信用等级为1的市政债券与CR为0的市政债券利率之差
- 信用等级的每一级有相同的偏效应，就可以通过3个约束条件来建立假设并进行检验
- 零假设 $\delta_2 = 2\delta_1, \delta_3 = 3\delta_1, \delta_4 = 4\delta_1$
- 限制模型为

$$MBR = \beta_0 + \delta_1(CR_1 + 2CR_2 + 3CR_3 + 4CR_4) + otherfactors$$

- 此时的系数就是前述的信用等级变量CR
- 顺序变量的取值太多时，通常采用分组的办法把相近的顺序变量作为一个组再设定虚拟变量进行回归分析

- 例4．4，2008-2010年本学院的金融硕士生毕业起薪与学生所来自的本科学校排名是否有关系？根据教育部985计划的分类，把学校分为三类，定义虚拟变量为：under9852为除北大清华外的前10名，under9853为排名更靠后的学校；北大和清华作为基准组。代入相关数据得到估计模型为

$$\log(\hat{salary}) = 10.741 - 0.374under9852 - 0.467under9853$$
$$(0.785)\quad (0.123)\quad (0.128)$$
$$+ 0.119y2008 - 0.264y2009 + 0.017econometric$$
$$(0.128)\quad (0.133)\quad (0.009)$$

- 如果采用2+8分组，得到估计模型为

$$\log(\hat{salary}) = 10.804 - 0.536under8 - 0.665under9 + 0.110y2008$$
$$(0.751)\quad (0.124)\quad (0.122)\quad (0.122)$$

- 4．1．5 虚拟变量与其它变量构成交叉项
- 把女性和已婚两个虚拟变量构成交叉项

$$\log(\hat{wage}) = 0.321 - 0.110\,female + 0.213married - 0.301\,female \cdot married + \cdots$$
$$(0.100)\quad (0.056)\quad (0.055)\quad (0.072)$$

- 可以检验不同性别的零假设而不依赖婚姻状况
- 也可以检验婚姻状况的差异而不依赖性别
- 允许不同组之间有不同的斜率系数
- 想检验在男性和女性之间受教育的回报是否一样

$$\log(wage) = \beta_0 + \delta_0\,female + (\beta_1 + \delta_1\,female)educ + \varepsilon$$

- 男性接受教育的收入增长为 β_1
- 女性接受教育的收入增长为 $\beta_1 + \delta_1$

- 要得到模型的OLS估计还需要把模型改写为

$$\log(wage) = \beta_0 + \delta_0\,female + \beta_1 educ + \delta_1\,female \cdot educ + \varepsilon$$

男女受教育的收益是否一样，零假设为 $H_0 : \delta_1 = 0$

- 接受同等教育水平的男女是否有相同的平均工资，零假设 $H_0 : \delta_0 = 0, \delta_1 = 0$

$$\log(\hat{wage}) = 0.389 - 0.227\,female + 0.082educ$$
$$(0.119)\quad (0.168)\quad (0.008)$$
$$- 0.0056\,female \cdot educ + 0.029\exp er - 0.00058\exp er^2$$
$$(0.0131)\quad (0.005)\quad (0.00011)$$
$$+ 0.032tenure - 0.00059tenure^2$$
$$(0.007)\quad (0.00024)$$

- 男性受教育收益的估计为8.2%，对女性的教育收益估计为0.082-0.0056=7.6%，它们之间的差为0.56%，t统计量为-0.0056/0.0131=-0.43
- 不论从经济意义还是统计意义都不显著
- F统计量得到F=14.69。在自由度为2和518时，这一F统计量非常大，p值非常小。
- 两种检验给出了完全不同的结果，但对这一问题的检验前面使用男女工资差为常数时的检验更合适

4．1．6 具有不同的回归方程的检验

检验两个组有相同的回归函数的零假设和两组之间有一个或多个斜率数是不一样的

上升市场和下降市场有相同的回归模型

$$return = \beta_0 + \beta_1 market + \beta_2 select + \beta_3 timig + \varepsilon$$

虚拟变量为市场收益为正时，up=1，市场收益为负时up=0

检验它们是否一样可以使用模型

$$return = \beta_0 + \delta_0 up + \beta_1 market + \delta_1 \cdot up \cdot market$$
$$+ \beta_2 \cdot select + \delta_2 \cdot up \cdot select + \beta_3 timig + \delta_3 \cdot up \cdot timig + \varepsilon$$

- 零假设为 $H_0:\delta_0=0,\delta_1=0,\delta_2=0,\delta_3=0$
- 计算F统计量还要估计去掉乘积项后的限制模型

$$return=\beta_0+\delta_0 up+\beta_1 market+\beta_2\cdot select+\beta_3 timig+\varepsilon$$

到回归模型的R方后，再计算F统计量

对k个不同自变量和截距项的模型进行检验

假定有两个组，分别称它们为g=1和g=2，模型为

$$y=\beta_{g,0}+\beta_{g,1}x_1+\cdots+\beta_{g,k}x_k+\varepsilon$$

零假设为上面方程中的每个 $\beta_{g,j}$ 都相同，共有k+1个约束条件，2(k+1)个参数

完全模型的残差平方和可以从两个模型中得到

$$SSR_\varepsilon=SSR_1+SSR_2$$

限制模型的残差平方和为把两组样本合起来估计一个方程得到 SSR_p

相应的F统计量为

$$F=\frac{[SSR_p-(SSR_1+SSR_2)]}{SSR_1+SSR_2}\cdot\frac{n-2(k+1)}{k+1},n=n_1+n_2$$

这一统计量的使用有一定的局限性

在同方差，而且两组样本的方差一样才合理

两个组之间的截距项不同，只检验它们之间斜率系数的差别，这时候要在对限制模型回归时加入一个分组的虚拟变量

- 例4．6，假若我们想检验中国证券市场CEO年薪在不同行业之间有不同的回归模型，制造业的回归模型与其他行业的回归模型不仅是截距项的不同，其他变量的系数可能也存在显著差异，模型为

$$\ln threeceo=\beta_0+\beta_1\ln asset+\beta_2 eachearn+\beta_3 first+\beta_4 second+\varepsilon$$

- 代入制造业的600家上市公司数据得到

$$\ln threeceo=9.08+0.196\ln asset+0.363 eachearn-0.579 first+1.091 second$$

（0.394） （0.018） （0.053） （0.230） （0.429）

N=600，R方0.234，残差平方和=358.9646

- 代入非制造业数据得到

$$\ln threeceo=7.441+0.280\ln asset+0.289 eachearn-0.194 first-0.174 second$$

（0.606） （0.029） （0.093） （0.229） （0.480）

n=454,R方=0.242，残差平方和=223.5756

$$F=\frac{[590.8061-(358.9646+223.5756)]}{358.9646+223.5756}\cdot\frac{1054-8)}{4}=3.71$$

4．1．7 二元因变量：线性概率模型

把虚拟变量作为因变量，用多元回归模型来解释一些定性事件

因变量y只取两个值：0和1

条件期望可以用离散分布的条件概率表示

$$E(y\mid x_1\cdots x_k)=\beta_0+\beta_1x_1+\cdots+\beta_kx_k$$

$$P(y=1\mid x_1,\cdots,x_k)=E(y\mid x_1,\cdots,x_k)$$

左边表示“成功”的概率，右边表示的是y=1的期望，为此我们可以把模型表示为

$$P(y=1\mid x_1\cdots x_k)=\beta_0+\beta_1x_1+\cdots+\beta_kx_k$$

$P(y=1\mid x_1\cdots x_k)$ 称为响应概率(response probability)

$$P(y=0\mid x_1\cdots x_k)=1-P(y=1\mid x_1\cdots x_k)$$

二元取值因变量的多元回归模型也称为线性概率模型（linear probability model ,LPM）

LPM中 β_j 度量的是其它因素不变时，x_j 的改变对成功概率的影响 $\Delta P(y=1\mid x_1\cdots x_k)=\beta_j\Delta x_j$

$\hat{\beta}_j$ 仍然表示了当 x_j 增加一个单位时对预计成功概率的影响

对因变量y=1时事件的名称正确表达非常重要

上市困境公司能否被ST之后两年内解脱困境

1998-2000年66家曾先后被ST公司的数据

$$\hat{succ}=14.6275+0.0144 l\arg est+0.0176 concen$$

（6.5776） （0.0123） （0.0075）

$$-1.7256Q-0.6081 size-5.3477 leverage$$

（0.6170） （0.2977） （2.0578）

股权集中度每增加一个百分点，就会导致成功摆脱困境的概率增加1.76%

托宾Q，公司规模和杠杆率都对摆脱困境有负的影响。

后两个变量的影响与直观和相关的金融理论相符。

可能是由于Q与杠杆率之间有较高的相关性

一些自变量组合给出的预测值不在0和1之间

概率不可能和所有的自变量都是线性

y的取值只有两个，LPM确实有不满足GM假设的问题

条件方差为

$$Var(y \mid x_1,\cdots,x_k) = P(x_1,\cdots,x_k)[1-P(x_1,\cdots,x_k)]$$

$P(x_1,\cdots x_k)$ 表示成功的概率

$$P(x_1,\cdots x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

除非这个概率 $P(x_1,\cdots x_k)$ 与任何一个自变量都没有关系，否则它一定是异方差的

4．1．8 策略分析

- 考虑贷款批准率的问题
- 不同种族的批准率不一样
- 批准率与许多因素有关，包括收入、财产、信用、还款能力

$$approved == \beta_0 + \beta_1 nonwhite + \beta_2 income + \beta_3 wealth + \beta_4 credrate + otherfactor$$

- 零假设 $H_0 : \beta_1 = 0$
- 非白人获得贷款与白人获得贷款的概率之差

4．2．异方差

- 使用t检验，F检验和置信区间对OLS估计进行推断时需要有同方差作保证
- 即使在大样本情形也需要
- 4．2．1 OLS中异方差的作用
- 同方差的假设5对OLS估计是否无偏或一致估计没有影响
- 评价模型的拟合度R^2和$\overline{R}^2$也不受异方差的影响
- 两个方差 $\sigma_\varepsilon^2, \sigma_y^2$都是真实方差而不是条件方差

- 存在异方差时，$Var(\hat{\beta}_j)$ 是有偏的
- LM统计量也不再渐近卡方分布
- 4．2．2 如果存在异方差可以找到比OLS更有效的估计
- 有幸的是，OLS估计仍然有用，只是需要对它进行适当的调整
- 称为异方差稳健方法 $Var(\hat{\beta}_j)$
- 存在异方差时如何估计
- 严格的理论推导很繁杂，但方法的应用却非常简捷

- 看一元回归的情形， $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- 假设GM假设1~4成立
- 存在异方差，则有 $Var(\varepsilon_i \mid x_i) = \sigma_i^2$
- OLS估计表示为

$$\hat{\beta}_1 = \beta_1 + [\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i] / \sum_{i=1}^n (x_i - \bar{x})^2$$

- 假设5不成立，可以推导出

$$Var(\hat{\beta}_1) = [\sum_{i=1}^n (x_i - \bar{x})^2 \sigma_i^2] / SST_x^2$$

- 如果 $\sigma_i^2 = \sigma^2$ 对所有I，上式为 σ^2 / SST_x

- 合理的估计方法不论是否存在异方差为

$$[\sum_{i=1}^n (x_i - \bar{x})^2 \hat{\varepsilon}_i^2] / SST_x^2$$

- 一般的多元回归模型

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

- 采用类似的方法得到一个无论是否存在异方差都合理的方差估计

$$\hat{Var}(\hat{\beta}_j) = [\sum_{i=1}^n \hat{r}_{ij}^2 \hat{\varepsilon}_i^2] / SSR_j^2$$

- 自由度调整的估计对上面的估计式乘n/(n-k-1)
- 构造异方差稳健的t统计量 $t = (\hat{\beta}_j - a_j) / \hat{se}(\hat{\beta}_j)$

- 例4．8，考虑异方差时CEO年薪与公司规模，盈利状况，股东分散度和销售收入的影响。采用前面同样的数据和模型重新估计异方差稳健的标准差，得到

$$\widehat{\log(CEO)} = 6.671 + 0.280\ln asset + 0.249 earn - 0.795 herfin + 0.0354\ln income$$

（0.411） （0.023） （0.046） （0.203） （0.011）
[0.411] [0.019] [0.055] [0.201] [0.012]

公司规模和盈利水平的OLS标准差与异方差稳健标准差明显有差别，规模的稳健异方差降低了大约20%，而盈利水平的稳健异方差提高了大约20%，其他变量的标准差变化不大。

- 例4.9 考虑异方差时工资和教育的影响

$$\widehat{\log(wage)} = 0.321 + 0.213 marrmale - 0.198 marrfem - 0.110 wingfem + 0.0789 educ$$

（0.100） （0.055） （0.058） （0.056） （0.0067）
[0.109] [0.057] [0.058] [0.057] [0.0074]

$$+ 0.0268\exp er - 0.00054\exp er^2 + 0.0291 tenure - 0.00053 tenure^2$$

(0.0055) (0.00011) (0.0068) (0.00023)
[0.0051] [0.00011] [0.0069] [0.00024]

- 几乎所有的异方差稳健标准差都比OLS标准差要大，但差别不是太大
- t统计量的显著性都没有改变
- 也有异方差稳健标准差小于OLS标准差的情况
- 异方差稳健标准差通常都会比OLS标准差大

- 既然异方差稳健标准差不论是否存在异方差都能给出合理的估计,为什么我们不直接使用它?
- 同方差的假设成立,t统计量不论样本量大小严格服从t分布,
- 异方差稳健标准差和稳健的t统计量只有在大样本才成立
- 对横截面数据具有大样本时，总是只给出异方差稳健标准差，也有两个都给出让读者比较的
- 同样可以给出在未知是否存在异方差和异方差形式时稳健的F统计量和LM统计量

- 4．2．3 计算异方差稳健的LM统计量
- 不需要特殊计量软件就能得到对检验排除多个约束时，异方差稳健的统计量
- 考虑如下的模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$$

- 要检验 $H_0 : \beta_4 = 0, \beta_5 = 0$
- 估计限制模型而得到残差 $\tilde{\varepsilon}$
- 再用所有的自变量对 $\tilde{\varepsilon}$ 进行回归得到 $R^2_{\tilde{\varepsilon}}$
- 常用的 LM统计量 $LM = n \cdot R^2_{\tilde{\varepsilon}}$
- 异方差稳健的LM还需要更多的步骤

- 用 x_1, x_2, x_3对x_4 回归得到的残差记为 $\tilde{r}_1$
- 用 x_1, x_2, x_3对x_5 回归得到的残差记为 $\tilde{r}_2$
- 用 $\tilde{r}_1\tilde{\varepsilon}, \tilde{r}_2\tilde{\varepsilon}$ 对1回归,不包含截距项
- 异方差稳健的LM统计量为 $n - SSR_1$
- SSR_1 为最后一个回归模型的残差平方和
- 在零假设下，LM渐近服从自由度为q的卡方分布
- 拒绝规则、p值的计算、置信区间等都与通常的LM统计量一样计算

- 例4．10，考虑在CEO年薪与公司规模，盈利状况和销售收入的影响模型中，使用异方差稳健的LM检验方法考察是否要加入第一和第二大股东持股

$$\hat{\varepsilon} = \widehat{\log(CEO)} - 6.944 - 0.263\ln asset - 0.229 earn - 0.034\ln income$$

$$\hat{r}_1 = firsthold + 0.282 - 0.025\ln asset - 0.032 earn - 0.004\ln income$$

$$\hat{r}_2 = \sec ondhold - 0.173 + 0.004\ln asset - 0.003 earn + 0.0003\ln income$$

n=1053 R^2 =0.0401 sum square=1033.26。

LM=1053-1033.26=19.74

查自由度为2的 χ^2_2 分布表得到1%的临界水平为9.21

4．2．4 对异方差的检验

- 在CLM假设之下，通常的t统计量严格服从t分布
- 存在异方差时，OLS不再是最优线性无偏估计
- 考虑下面的模型

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k+\varepsilon$$

- 假设1~4成立，关注的零假设为假设5为真

$$H_0:Var(\varepsilon\mid x_1,\cdots,x_k)=\sigma^2$$

- 不能拒绝零假设通常就认为异方差的问题不大
- 干扰项条件期望为0 $Var(\varepsilon\mid x)=E(\varepsilon^2\mid x)$
- 同方差的假设也等价于 $H_0:E(\varepsilon^2\mid x_1,\cdots,x_k)=\sigma^2$

- 4．2．5 异方差的Goldfelt-Quandt检验
- 把样本按这一变量进行分组
- 估计两个OLS回归线：一根是低残差方差部分，一根是高残差方差部分
- 检验方法可以通过下面的步骤实现：

1)用怀疑存在异方差的自变量把样本数据排序；

2)删除比例为d/n的中间样本，d是选取的数量

3)分别使用自变量取值低的部分和高的部分，回归模型的样本量都为(n-d)/2

4)计算每一回归模型的残差平方和，自变量取值低的部分的残差平方和为 SSR_1，自变量取值高的部分的残差平方和为 SSR_2

5）假定残差序列服从正态分布，且是相互独立的，则检验统计量：$F=SSR_2/SSR_1$

给出原模型是否存在异方差的拒绝规则

检验方法存在一定的不足

没有对估计参数进行约束，检验的功效有所损失。功效更强的检验是限制两个部分回归的参数是相同的，只是方差不等

删除中间部分样本的比例带有一定的随意性

- 例4．11，考虑在CEO年薪与公司规模，盈利状况的影响模型中，假若我们怀疑公司的盈利水平高低导致CEO的年薪变化方差不同，考虑使用Goldfelt-Quandt检验方法来考察是否存在异方差
- 用公司盈利水平最低的362家公司给出估计模型为

$$\widehat{\log(CEO)}=7.057+0.287\ln asset+0.141earn-0.425\,first+0.838\sec ond$$

Sum=174.69，公司盈利水平最高的362家公司模型为

$$\widehat{\log(CEO)}=7.057+0.287\ln asset+0.141earn-0.425\,first+0.838\sec ond$$

Sum=207.87， 根据两个回归模型的残差平方和计算F检验统计量为： F=207.87/174.69=1.19

自由度为357,357的F分布在5%的临界水平为1.19

4．2．6 Brensch-Pagan异方差检验

- 实际上是检验 ε^2是否与一个或多个自变量有关
- 最简单的方法就是假设是线性函数

$$\varepsilon^2=\delta_0+\delta_1x_1+\delta_2x_2+\cdots\delta_kx_k+u$$

- 同方差的假设 $H_0:\delta_1=\delta_2=\cdots=\delta_k=0$
- 对整个模型解释能力进行检验的F或LM统计量来对这些自变量全体对是否有整体显著性的检验
- 不知道模型的真实误差项，使用OLS估计

$$\hat{\varepsilon}^2=\delta_0+\delta_1r_1+\cdots+\delta_kr_k+v$$

- 可证明OLS估计的残差代替真实的干扰项不会影响F或LM统计量的大样本分布

- 检验的F统计量为 $F=\dfrac{[R^2_{\hat{\varepsilon}^2}/k]}{[(1-R^2_{\hat{\varepsilon}^2})/(n-k-1)]}$
- LM统计量 $LM=n\cdot R^2_{\hat{\varepsilon}^2}$
- 检验是否存在异方差时，可以总结为以下步骤

(1)估计通常回归模型得到OLS回归残差 $\hat{\varepsilon}_i^2,i=1,\cdots,n$

(2)估计回归模型 $\hat{\varepsilon}^2=\delta_0+\delta_1x_1+\cdots+\beta_k\delta_k+v$ 得到OLS估计下的 $R^2_{\hat{\varepsilon}^2}$

(3)构造F和LM统计量渐近分布分别为 $F_{k,n-k-1}$ χ^2_k

- 若怀疑异方差主要是依赖于某些自变量，只需要用怀疑的这些自变量对 $\hat{\varepsilon}^2$ 进行回归

- 例4．12，考虑在CEO年薪与公司规模，盈利状况的影响模型中，假若我们怀疑公司的盈利水平高低导致CEO的年薪变化方差不同，考虑使用Goldfelt-Quandt检验方法来考察是否存在异方差

$$\widehat{\log(CEO)} = 6.572 + 0.319\ln asset + 0.252 earn - 0.566\, first + 0.650\sec ond$$

用上一模型得到的残差序列平方值为因变量给出模型为

$$\hat{\varepsilon}^2 = 1.590 - 0.055\ln asset + 0.219 earn + 0.161\, first + 0.072\sec ond$$

Sum=880.60　n=1053，R方=0.016

根据两个回归模型的残差平方和计算F检验统计量

$$F = \frac{[R^2_{\hat{\varepsilon}^2}/k]}{[(1-R^2_{\hat{\varepsilon}^2})/(n-k-1)]} = \frac{[0.016/4]}{[(1-0.016)/(1053-4-1)]} = 4.26$$

自由度为4,1048的F分布表得到在1%的临界水平为3.34

4．2．7 异方差的White检验

- GM假设成立时，通常的OLS标准差和检验统计量是渐近合理的，同方差的假设可以换为更弱的假设 ε^2和所有的自变量都不相关，与所有自变量平方 x_j^2 及自变量间的交叉项 $x_j x_{h,\,j+h}$ 都不相关
- 模型中有6个自变量，则White检验的回归中将涉及到27个自变量
- 考虑使用自变量的一些函数形式
- 每个拟合值为 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}$
- 正好是自变量的一个线性函数

- 对它平方就得到自变量的各个平方项和交叉项
- 可用模型 $\hat{\varepsilon}^2 = \delta_0 + \delta_1\hat{y} + \delta_2\hat{y}^2 + v$ 来检验异方差
- 可用F和LM统计量来检验零假设 $H_0 : \delta_1 = 0, \delta_2 = 0$
- 一个非常好的转换，而且大大简化了检验的过程
- 寻找存在异方差的证据来拒绝同方差的假设
- 假设3不成立时，E（y|x）的函数形式设定不合适，则即使E（y|x）是常数也会得到拒绝的结论
- 模型需要使用对数而使用了水平时，异方差的检验就会显著
- 回归方程形式的设定不当比存在异方差带来的危害要更大

- 例4．13，考虑在CEO年薪与公司规模，盈利状况的影响模型中，假若我们怀疑公司的盈利水平高低导致CEO的年薪变化方差不同，考虑使用White检验方法来考察是否存在异方差。

$$\widehat{\log(CEO)} = 6.572 + 0.319\ln asset + 0.252 earn - 0.566\, first + 0.650\sec ond$$

$$\widehat{\log(CEO)} = 6.843 + 0.3\ln asset + 0.236 earn$$

用拟合值的一次项和二次项，只使用拟合值的一次项分别对残差序列平方值作为因变量给出估计模型为

$$\hat{\varepsilon}^2 = 0.133\widehat{\log(CEO)} - 0.007\log(CEO)\text{^}2$$

（0.407）　（0.019）

$$\hat{\varepsilon}^2 = 0.041\widehat{\log(CEO)}$$

（0.002）

4.2．8 加权最小二乘估计

- A）已知异方差的形式
- 异方差的形式为 $Var(\varepsilon \mid x) = \sigma^2 h(x)$
- 需要h(x)对所有自变量的取值都取正值
- 随机抽取的一组样本，对应的方差表示为

$$\sigma_i^2 = Var(\varepsilon_i \mid x_i) = \sigma^2 h(x_i) = \sigma^2 h_i$$

- 简单的存款模型中

$$sav_i = \beta_0 + \beta_1 inc_i + \varepsilon_i, Var(\varepsilon_i \mid inc_i) = \sigma^2 inc_i$$

- h(x)=inc，即干扰项的方差与收入水平成比例
- 是否可以利用条件方差来给出 β_i 的估计呢？

- 要考虑的模型为

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

- 干扰项包含异方差的函数形式h(x)是已知的
- 把这一模型转化为满足GM假设的同方差模型
- 已知 x_i 时，h_i 也是已知的，所以 $\varepsilon_i / \sqrt{h_i}$ 的期望为0；进一步还有 $Var(\varepsilon_i \mid x_i) = E(\varepsilon_i^2 \mid x_i) = \sigma^2 h_i$

$$E[(\varepsilon_i/\sqrt{h_i})^2] = E[\varepsilon_i^2]/h_i = (\sigma^2 h_i)/h_i = \sigma^2$$

- 原来模型两边除以 $\sqrt{h_i}$ 得到

$$y_i/\sqrt{h_i} = \beta_0/\sqrt{h_i} + \beta_1(x_{i1}/\sqrt{h_i}) + \cdots + \beta_k(x_{ik}\sqrt{h_i}) + \varepsilon_i/\sqrt{h_i}$$

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

- 只有截距项的系数改变，其它的参数仍然一样
- 表达模型参数的意义时，需要返回原来的模型
- 存款的例子中，变换后的模型为

$$sav_i/\sqrt{inc_i}=\beta_0(1/\sqrt{inc_i})+\beta_1\sqrt{inc_i}+\varepsilon_i^*$$

- 如果原来的模型满足GM假设1~4，则变换后的模型满足GM假设1~5
- 变换后模型用OLS估计得到 $\beta_0^*,\beta_1^*,\cdots,\beta_k^*$ 将与从原来模型用OLS估计得到的 $\hat{\beta}_1,\cdots,\hat{\beta}_k$ 不同
- β_j^*称为广义最小二乘估计（generalized least squares, GLS）

- 变换后模型得到的标准差估计，t统计量、F统计量都具有理想的性质
- β_j^* 是最优线性无偏估计，比 $\hat{\beta}_j$ 更有效
- 变换后的模型得到更有效的估计
- 变化后模型的R方也没有了度量拟合度的功能
- 应用中很难知道方差依赖于自变量的函数形式
- 但在有些特殊的数据类型中，存在一种天然的异方差结构
- 不是使用每个具体单元的权重，而只是对每一组的平均水平作为该组的权重

- 例4．16，员工对401（k）缴存额和公司匹配率之间的关系
- 假定i为给定的公司，e为给定公司内的每一位员工

$$contrib_{i,e}=\beta_0+\beta_1 earns_{i,e}+\beta_2 age_{i,e}+\beta_3 mrate_i+\varepsilon_{i,e}$$

- 假若没有得到每个员工的缴存额、收入和年龄，只有每家公司的平均数
- 则对公司的平均水平可以得到模型

$$\overline{contrib}_{i,e}=\beta_0+\beta_1\overline{earns}_{i,e}+\beta_2\overline{age}_{i,e}+\beta_3\overline{mrate}_i+\bar{\varepsilon}_i$$

- 原模型满足GM假设，每个人的误差与公司规模独立
- 当原模型满足同方差的假设时，则公司层面的模型必然是异方差的

$$Var(\varepsilon_{i,e})=\sigma^2 \qquad Var(\bar{\varepsilon}_i)=\sigma^2/m_i$$

- 最有效的估计就是WLS，每家公司的权重正好是其员工数 $1/h_i=m_i$
- 大公司给予更多的权重
- 类推到对每个城市、国家和地区水平
- 使用WLS方法并同时给出其稳健的标准差，确保给出的估计是有效的

b). 纠正异方差的GLS估计
纠正异方差的GLS估计也称为加权最小二乘（weighted least squares, WLS）估计

- 最小化一个加权的残差平方和来得到 β_j^*
- 求和项的权重为 $1/h_i$

$$\sum_{i=1}^{n}(y_i-b_0-b_1x_{i1}-\cdots-b_kx_{ik})^2/h_i$$

- 把权重因子放入平方项

$$\sum_{i=1}^{n}(y^*_i-b_0x^*_{i0}-b_1x^*_{i1}-\cdots-b_ix^*_{ik})^2$$

- WLS估计可通过定义任何一组为正的权重来获得

- 例4．15，在硕士生年薪的例子中，得到了2007-2009三年毕业的179位同学的年薪，本科成绩，本科院校，计量成绩，工作单位性质等数据，分别使用OLS和WLS并考虑前面所说的异方差形式
- 分别采用OLS和WLS对年薪与学习成绩的几种估计模型见表4.1。
- 分别使用white 和BP检验的方法构建异方差权重，考虑拟合值平方项影响后，计量成绩和GPA的影响都变得不再显著；
- 使用BP方法，只考虑变量的线性项对异方差的影响，给出的模型提高了计量和GPA的影响

用OLS和WLS估计的成绩与收入模型

自变量	OLS	WLS(ec)	OLS	WLS（whit	WLS(BP)
Inc	10.80 (0.751)	10.67 (0.751)	9.967 (0.931)	10.333 (0.948)	9.586 (0.964)
Y2008	0.110 (0.122)	0.101 (0.122)	0.055 (0.122)	0.068 (0.123)	0.090 (0.117)
Y2009	-0.296 (0.127)	-0.304 (0.127)	-0.353 (0.127)	-0.333 (0.120)	-0.327 (0.126)
性别			0.035 (0.111)	0.045 (0.108)	0.057 (0.106)
计量成绩	0.018 (0.009)	0.020 (0.008)	0.015 (0.009)	0.013 (0.009)	0.016 (0.009)
GPA			0.343 (0.213)	0.283 (0.211)	0.414 (0.225)
本科院校8	-0.536 (0.124)	-0.531 (0.123)	-0.656 (0.132)	-0.651 (0.139)	-0.643 (0.138)
本科院校其他	-0.665 (0.122)	-0.656 (0.121)	-0.726 (0.128)	-0.697 (0.135)	-0.735 (0.137)
样本量	159	159	153	153	153
R方	0.223		0.241		

4．2．9 估计异方差的函数形式

- 应用中异方差的函数形式很少会知道
- 对h进行建模，通过数据来给出它的参数估计
- 介绍一种比较简便的可行方法
- 假设 $Var(\varepsilon \mid x) = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k)$
- 则有 $h(x) = \exp(\delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k)$
- 目前的任务是如何得到 δ_j 的估计
- 假定干扰 v与自变量独立得到

$$\log(\varepsilon^2) = \alpha_0 + \delta_1 x_1 + \cdots + \delta_k x_k + u$$

- 实际上是用 $x_1, \cdots, x_k$ 对 $\log(\hat{\varepsilon}^2)$ 进行回归

- 使用的是这一回归模型的拟合值 $\hat{g}_i$

$$h_i\text{的估计为}\hat{h}_i = \exp(\hat{g}_i)$$

- 调整异方差的FGLS估计方法为：
- 1）用 $x_1, \cdots, x_k$对y回归得到残差；
- 2）计算 $\log(\hat{\varepsilon}^2)$
- 3）用 $x_1, \cdots, x_k$对 $\log(\hat{\varepsilon}^2)$ 进行回归得到每一个观测值对应的模型拟合值 $\hat{g}_i$
- 4）计算 $\hat{h} = \exp(\hat{g}_i)$；
- 5）用WLS方法估计模型 $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$，权重为 $1/\hat{h}$

- 采用 $\hat{h}_i$ 为权重得到的估计不再是最优线性无偏估计，但它是一致的，且比OLS更渐近有效
- 在大样本时 FGLS比OLS更好
- 可用OLS的拟合值及其平方作自变量来得到 $\hat{g}_i$
- 即 $\hat{g}_i$ 为用 $\hat{y}, \hat{y}^2$对$\log(\hat{\varepsilon}^2)$ 进行回归得到的拟合值
- 用WLS估计的结果来进行F统计量计算，要特别注意在完全模型和限制模型中使用完全相同的权重
- 应用实例中，OLS和WLS给出的估计会有明显的差异，有时甚至改变模型的结果
- 干扰项与任一自变量之间存在相关性时，就会引起OLS和WLS之间的不一致

- 例4．16，在硕士生年薪的例子中，得到了2007-2009三年毕业的179位同学的年薪，本科成绩，本科院校，计量成绩，工作单位性质等数据，分别使用OLS和WLS并考虑前面所说的异方差形式。

$$\hat{\varepsilon}_i = \ln salary - 9.586 - 0.09\,y2008 + 0.327\,y2009 - 0.016\,econ + 0.643\,und8$$
（0.964） （0.117） （0.126） （0.009） （0.138）
$$+ 0.735\,und9 - 0.057\,man - 0.414\,GPA$$
（0.137） （0.106） （0.225）

3)．用 $x_1, \cdots, x_k$ 对$\log(\hat{\varepsilon}^2)$ 进行回归得对应的模型拟合值

$$\hat{g}_i = -8.217 - 0.312\,y2008 - 0.568\,y2009 + 0.040\,econ - 1.33\,und8$$
（3.70） （0.448） （0.483） （0.034） （0.529）
$$+ 1.495\,und9 + 0.164\,man + 1.072\,GPA$$
（0.526） （0.406） （0.863）

4)．计算 $\hat{h}_i = \exp(\hat{g}_i)$

- 5)．用WLS方法估计模型

$$\ln salary = 10.283 + 0.037\,y2008 - 0.347\,y2009 + 0.019\,econ - 0.578\,under8$$
（0.877） （0.117） （0.120） （0.0086） （0.156）
$$- 0.602\,under9 + 0.086\,man + 0.124\,GPA$$
（0.154） （0.103） （0.185）

- 使用WLS估计的结果进行F统计量的计算，我们要特别注意在完全模型和限制模型中要使用完全相同的权重。
- 应用中，OLS估计和WLS估计给出的结果会有明显的差异，有时甚至改变模型的结果
- 如果OLS和WLS估计给出的结果有符号差异，并且结果显著或估计系数相差特别大，就要特别小心。这时GM假设也许不成立，

4．2．10 线性概率模型的异方差

- 模型必然包含异方差
- 简单地使用OLS估计不能给出LPM的有效估计
- LPM的条件方差为：Var(y|x)=p(x)[1-p(x)]

$$p(x)=\beta_0+\beta_1x_1+\cdots+\beta_kx_k$$

- p(x)显然依赖于未知的真实参数
- 对每一个观测值i有$\mathrm{Var}(y_i\mid x_i)$ 的估计 $\hat{h}_i=\hat{y}_i(1-\hat{y}_i)$
- 在此所讨论的 $\hat{y}_i$ 是一个概率，它必须在0，1之间
- 如果大部分的观测值都满足$0<\hat{y}_i<1$ 通常的办法是调整这些不满足的取值

55

- 对 $\hat{y}_i\le 0$取$\hat{y}_i=0.01$
- 对 $\hat{y}_i\ge 1$取$\hat{y}_i=0.99$或其它的值
- 用WLS方法估计LMP的步骤：
- 1）用OLS方法估计模型得到拟合值；
- 2）确定是否所有的拟合值都在（0，1）区间内，如果是就进入第三步，如果不是就通过某种方法把它们调整到（0，1）区间内；
- 3）计算异方差 $\hat{h}_i=\hat{y}_i(1-\hat{y}_i)$
- 4）使用权重系数 $1/\hat{h}$ 用WLS方法估计模型

$$y=\beta_0+\beta_1x_1+\cdots+\beta_kx_k+\varepsilon$$

56

4．3．模型的多重共线性问题

- 假设4并不是不允许自变量之间存在相关性，而只是限制不能完全相关
- 当一个或几个自变量之间具有比较高的相关性时，称为多重共线性
- 多重共线性是R_j^2 比较接近1或XX'虽可逆，由于其部分取值比较小，使逆矩阵 $(XX')^{-1}$ 的数值非常大
  - 没有一个明确的意义和界限

57

4．3．1 多重共线性的影响

- OLS估计仍然是无偏的，但由于估计的标准差比较大，很难给出估计系数偏效应的合理估计，估计可以偏离真实值很远
- 使得t统计量不容易达到显著水平，也很难给出对估计系数的推断
- 容易高估(overestimates)或低估(underestimates)该变量所具有的真实偏效应
- 是否有一些系数的估计标准差比较大
- 正规的处理方法是考察自变量的方差-协方差矩阵

58

4．3．2 多重共线性的判断

- 常用的判别共线性的方法
  - 数据很小的变化可能带来估计系数比较大的变动
  - 模型的系数具有非预期的符号，或者是有不成比例的系数
  - 一个模型有相对比较高的R方，但具体变量的t统计量显著程度不高
  - 自变量两两之间的相关系数
  - 条件数(condition number)是矩阵的最大特征根与最小特征根的比值的平方根

59

$$\gamma=\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$$

- 条件数大于20或30被认为是存在共线性的特征
- 把矩阵进行方差分解来判断具体某个自变量的共线性程度的方差因子（variance inflation factors (VIF)）或单一变量的容忍度（tolerances for individual variables）
- 容忍度就是 $1-R_j^2$
- VIF就是 $1/(1-R_j^2)$

60

- 例4．18，在CEO年薪的例子中，采用2007年的数据，只考虑公司的盈利和股东持股情况回归模型为

$$\log(threeceo) = 2.835 + 0.292\ln asset + 0.029\ln income + 0.106 eachearn - 0.231 firsthold$$
（0.181） （0.024） （0.012） （0.020） （0.072）
$$+ 0.190\sec ondhold + 0.566 thirdhold$$
（0.148） (0.378)

各变量的VIF分别为：1.73，1.61，1.26，1.26，1.37
最大的特征为：0.809，最小的特征根为0.001；
由此可以计算出模型的条件系数为65.896。模型存在比较严重的共线性问题

4．3．3 多重共线性的处理方式

- 简单直接方法是直接从模型中删除产生问题的变量
- 删除变量涉及到模型设定的问题
- 寻找稍微有偏，但估计的方差更小的方法
- 两种针对可能存在共线性时标准差太大的处理方法
- 岭回归估计方法在OLS估计上再加一个对角阵

$$\beta_r = (X'X + rD)^{-1}X'Y$$

- 岭回归估计是有偏的，但可以验证其估计的方差

$$Var(\beta_r) = \sigma^2(X'X + rD)^{-1}X'X(X'X + rD)^{-1}$$

- 比OLS估计的方差小

- 主成份方法：
- 通过所有自变量的方差协方差矩阵分解而得到矩阵的特征根，特征向量，使用最大的几个特征根对应的特征向量作为新的回归模型的自变量而给出模型的估计。

4．4．有关模型设定和数据的问题

- 干扰项与其中一个或多个自变量-有关内生的变量
- 函数形式设定不当
  - 函数形式设定不当的检验方法RESET
  - 对不可观测的自变量使用代理变量
  - 可以使用滞后的因变量作为代理变量
  - 在变量的度量有误差时OLS估计的性质
  - 在自变量中存在度量误差
  - 数据缺失，非随即抽样和异常值的处理
  - 异常值的影响

4．4．1 函数形式设定不当

- 建立模型时没有正确考虑自变量和因变量间关系
- 例如，在小时工资的例子中模型为

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 \exp er + \beta_3 \exp er^2 + \varepsilon$$

- 工龄的平方项没有包含在模型中
- 将会导致估计 $\hat{\beta}_0, \hat{\beta}_1$和$\hat{\beta}_2$有偏
- 因为工龄与收益之间的关系为 $\beta_2 + 2\beta_3 \exp er$
- 设定不当模型中，这个影响只有 β_2

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 \exp er + \beta_3 \exp er^2 + \beta_4 female + \beta_5 female \cdot educ + \varepsilon$$

- 模型中去掉交叉项也面临函数形式设定不当问题

- 缺失变量不是函数形式设定不当的唯一原因
- 不是使用log(wage)，而直接使用wage作为因变量，也无法得到各自变量偏效应的无偏或一致估计
- F统计量可以用来检验函数形式设定不当的问题
- 但加入显著性的平方项也会带来其它的渐近函数形式问题
- 本来是应该采用对数的模型，使用水平再加上平方项就会给出一个近似

4．4．2 函数形式设定不当的检验方法RESET

- Ramsey(1969)提出的回归设定误差检验
- RESET通过OLS拟合值的多项式来检查函数形式设定是否适当
- 通常是考虑二次和三次项
- 考虑一个加项的模型

$$y=\beta_0+\beta_1x_1+\cdots+\beta_kx_k+\delta_1\hat{y}^2+\delta_2\hat{y}^3+u$$

- 零假设为初始模型设定适当 $H_0:\delta_1=0,\delta_2=0$
- F统计量显著说明模型有函数形式设定不当的问题

- 例4．19，在硕士起薪例子中，154位同学，分别使用拟合值来检验是否需要加入平方项和交叉项

$$\ln salary=9.383+0.088y2008-0.306y2009-0.684under8-0.746under9$$
$$(1.009)\quad(0.122)\quad(0.127)\quad(0.132)\quad(0.127)$$
$$+0.044agend+0.597GPA+0.01econometric$$
$$(0.110)\quad(0.244)\quad(0.009)$$
N=177，R=0.275， adjR=0.24， sum=50.4046

$$\ln salary=20.325+0.384y2008-1.314y2009-2.971under8-3.233under9$$
$$(20.071)\quad(0.557)\quad(1.850)\quad(4.191)\quad(4.559)$$
$$+1.484agend+2.588GPA+0.045econometric-0.136\ln \hat{salary}\text{^}2$$
$$(2.641)\quad(3.655)\quad(0.064)\quad(0.250)$$
N=177， R=0.277， adjR=0.236， sum=50.2998

$$F=\frac{[SSR_r-SSR_u)]}{SSR_u}\cdot\frac{n-2(k+1)}{q}=\frac{[50.4046-50.2998)]}{50.2998}\cdot\frac{177-(8+1)}{1}=0.35$$

- 可以给出相应的LM检验，甚至还可以采用前一节的方法给出异方差稳健的检验
- 只能给出不适当的检验结果，并不能给出如何改进模型的方向
- 只是一个检验函数形式设定不当的工具
- 另一类检验函数形式设定不当的方法是考察自变量是应该使用水平还是使用对数

$y=\beta_0+\beta_1x_1+\beta_2x_2+\varepsilon$和$y=\beta_0+\beta_1\log(x_1)+\beta_2\log(x_2)+\varepsilon$

- 不能直接使用标准的F检验
- 构造一个包含全部自变量的完全模型

$$y=\gamma_0+\gamma_1x_1+\gamma_2x_2+\gamma_3\log(x_1)+\gamma_4\log(x_4)+u$$

- 零假设 $H_0:\gamma_3=0,\gamma_4=0$ 或 $H_0:\gamma_1=0,\gamma_2=0$
- 第二种检验方法是基于如果前一模型是对的，则后一模型得到的拟合值在前一模型中应该不显著
- 用t统计量检验模型中的系数 $\hat{\hat{y}}$

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\theta_1\hat{\hat{y}}+u$$

- 这类检验也存在一些问题
- 首先不一定有一个明确的结果，可能两个模型同时被拒绝，也可能两个都不能拒绝
- 其次拒绝了任何一个并不表示另一个模型就是对的

- 例4．20，在CEO年薪例子中，检验是水平变量合适还是对数形式合适。
- 使用所有的水平变量和对数变量得到

$$\ln threeceo=6.243+0.296\ln asset+0.036\ln income-7.90asset+0.228income$$
$$(0.426)\quad(0.024)\quad(0.012)\quad(11.00)\quad(0.906)$$

分别使用水平和对数变量得到拟合值

$$\ln threeceo=7.070+1.68asset+0.644income+0.311\hat{\hat{y}}$$
$$(0.435)\quad(1.11)\quad(0.936)\quad(0.022)$$
$$\ln threeceo=8.823+0.296\ln asset+0.036\ln income-0.194\hat{y}$$
$$(2.449)\quad(0.024)\quad(0.012)\quad(0.193)$$

使用对数的拟合值代入水平变量模型中t统计量为14.14
用水平值的拟合值代入对数模型t统计量为1.00

4．4．3 对不可观测的自变量使用代理变量

- 最难处理的问题是由于数据缺失而没有在模型中包含关键的变量
- 工资与教育的例子中假定这一问题的真实模型为

$$\log(wage)=\beta_0+\beta_1educ+\beta_2\exp er+\beta_3abil+\varepsilon$$

- 因为能力不能被测到，如何才能解决这一问题，或者从一定程度上减缓由于缺失变量带来的偏差
- 对缺失变量使用代理变量（proxy variable）
- 先天能力进行替代的可能变量就是智商或IQ
- 代理变量也只需要与原来的变量有关即可

- 考虑模型为 $y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3^*+\varepsilon$
- 假定y,x_1,x_2的数据可以得到，解释变量x_3^*是不可观测的，但它有一代理变量x_3
- 对代理变量基本的要求是它与 x_3^* 有关系

$$x_3^*=\delta_0+\delta_3x_3+u_3$$

- 通常可以认为x_3^*和 x_3之间是正相关的
- 截距可正可负，容许x_3^*和 x_3间有不同的度量尺度
- 实际中就是认为x_3^*和 x_3 是一样的来实施回归
- 称为缺失变量问题的插入式解

- 要使β_1和 β_2 的插入式解为一致估计需要对 ε u_3作出假设
- 首先需要干扰项 ε和 x_1,x_2,x_3^*不相关，此外还要求ε 和 x_3 不相关
- u_3与 x_1,x_2,x_3不相关
- 用条件期望的表达

$$E(x_3^*\mid x_1,x_2,x_3)=E(x_3^*\mid x_3)=\delta_0+\delta_3x_3$$

- 考虑了 x_3 的偏效应之后， x_1,x_2与x_3^*不相关
- 人的先天能力的平均水平只会随IQ而改变，不会随着受教育程度和工龄增加或升职有关

- 考虑在这些假设下插入式解的结果

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3(\delta_0+\delta_3x_3+u_3)+\varepsilon$$
$$=(\beta_0+\beta_3\delta_0)+\beta_1x_1+\beta_2x_2+\beta_3\delta_3x_3+\varepsilon+\beta_3u_3$$

- 干扰项组合 $e=\varepsilon+\beta_3u_3$ 依赖于初始模型和代理变量模型干扰项，期望是零，而且与 x_1,x_2,x_3 无关

$$y=\alpha_0+\beta_1x_1+\beta_2x_2+\alpha_3x_3+e$$

- $\alpha_0=\beta_0+\beta_3\delta_0$ 是新模型的截距项，$\alpha_3=\beta_3\delta_3$是代理变量 x_3 的斜率参数
- 对新模型使用OLS估计可以得到 α_0,β_1,β_2和α_3 的无偏估计

- 用IQ作为能力的代理变量，对935个人的调查，有每月的收入、教育程度和其它一些特征变量
- 第一列是没有包含代理变量IQ的估计结果，估计的教育回报为6.5%
- 加入了IQ之后，教育回报下降到5.4%
- 其它条件相同时，IQ增加10个点预计会增加3.6%的每月收入
- 考虑了IQ、教育等之后，白人和黑人之间仍然有14.3%的收入差

自变量	(1)	(2)	(3)	自变量	(1)	(2)	(3)
Educ	0.065 (0.006)	0.054 (0.007)	0.018 (0.041)	Urban	0.184 (0.027)	0.182 (0.027)	0.184 (0.027)
Exper	0.014 (0.003)	0.014 (0.003)	0.014 (0.003)	Black	-0.188 (0.038)	-0.143 (0.039)	-0.147 (0.040)
Tenure	0.012 (0.002)	0.011 (0.002)	0.011 (0.002)	IQ		0.0036 (0.0010)	-0.0009 (0.0052)
Married	0.199 (0.039)	0.200 (0.039)	0.201 (0.039)	Educ*IQ			0.00034 (0.00038)
South	-0.091 (0.026)	-0.080 (0.026)	-0.080 (0.026)	Intercept	5.395 (0.113)	5.176 (0.128)	5.648 (0.546)
Observation	935	935	935	R-Squared	0.253	0.263	0.263

- 假定不可观测变量不仅与代理变量相关，而且还与其它的自变量有关

$$x_3^*=\delta_0+\delta_1x_1+\delta_2x_2+\delta_3x_3+u_3$$

- 前面假设 δ_1和δ_2都为0。方程插入初始模型得到

$$y=(\beta_0+\beta_3\delta_0)+(\beta_1+\beta_3\delta_1)x_1+(\beta_2+\beta_3\delta_2)x_2+\beta_3\delta_3x_3+\varepsilon+\beta_3u_3$$

$$p\lim(\hat{\beta}_1)=\beta_1+\beta_3\delta_1,\ p\lim(\hat{\beta}_2)=\beta_2+\beta_3\delta_2$$

- 如果能力和教育的偏相关为正，则估计有正偏差
- 若IQ不是好的代理变量，所得估计将会是向上偏
- 但我们有理由相信这个偏离会比缺失能力变量时的偏差要小

4．4．4 使用滞后的因变量作为代理变量

- 不知道如何去寻找代理变量
- 用因变量过去时期的取值
- CEO年薪例中一些公司一直就可能付比较高工资
- 许多不可观测到的因素同时会影响到过去的和现在的年薪，解释的模型

$$salary = \beta_0 + \beta_1 ROE + \beta_2 ind + \beta_3 salary_{-1} + \varepsilon$$

- 预计 $\beta_3 > 0$
- 至少看到两个公司过去有相同收入和相同行业
- 使用面板数据能给出更多的处理方式

- 例4．21，在CEO年薪的例子中，如果只考虑公司的规模、盈利能力，大股东的分散程度，我们可以得到回归模型为

$$\ln threeceo = 8.355 + 0.198\ln asset + 0.037\ln income + 0.337 eachearn - 0.609 herfindahl\_5$$
（0.335）（0.017） （0.010） （0.046） （0.206）
n=1051, R^2=0.229

用上一年的CEO年薪来反映一些公司潜在的不可观测因素影响，得到估计模型为：

$$\ln ceo = 2.392 + 0.048\ln asset + 0.016\ln income + 0.145 earn - 0.093 herfind + 0.729 CEO_{-1}$$
（0.282）（0.012） （0.007） （0.031） （0.140） (0.021)
n=1017, R方=0.657

控制了上一年的CEO年薪后，其他变量的系数都出现了比较大的下降，特别是大股东的分散程度影响变得很弱了，而模型的解释程度大幅的提高了，这一定程度说明我们前面模型中所考虑的几个变量只是影响CEO年薪的一部分因素，还有很多因素没有在模型中得到考虑

4．4．5 变量的度量有误差时OLS估计的性质

- 存在度量误差时如何给出OLS估计
- 统计结构上比较类似于缺失变量而使用代理变量的问题，但在概念上是不一样的
- 不是不可观测变量，只是所记录的数据有误差
- 度量误差问题中的自变量通常就是所关注的变量

4．4．6 因变量有度量误差问题

- 一个家庭的年度存款额，回归模型为

$$y^* = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

- 调查的或报告出来的存款额通常是不准确的

- 度量的误差定义为 $e_0 = y - y^*$
- 度量误差怎么与其它因素关联
- 把误差方程插入到原来的模型中

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon + e_0$$

- 什么情况下用y代替y*时能给出一致估计？
- 因原来的模型满足GM假设
- 对 e_0 首先是期望为0的假设
- 最重要是如何假设度量误差 ε与x_j 之间的关系
- 直接假设是y中的度量误差与每一个自变量都独立

- 条件成立时，给出的OLS估计将是无偏的和一致的。且有通常推断方法中的t，F和LM统计量都是合理的
- 如果还假设 e_0 与 ε 不相关 ，则

$$Var(\varepsilon + e_0) = \sigma_\varepsilon^2 + \sigma_0^2 > \sigma_\varepsilon^2$$

- 因变量中有度量误差时导致估计的方差增大
- 保证OLS估计有较好性质的条件是度量误差与自变量不相关
- 例如，存款函数具有度量误差

$$sav^* = \beta_0 + \beta_1 inc + \beta_2 size + \beta_3 age + \varepsilon$$

- 假设度量误差与inc，size，educ和age无关
- 高收入的家庭和接受教育程度高的人可能报告的存款更不准确
- 因变量使用对数形式时，度量误差表现为乘积形式

$$\log(y) = \log(y^*) + e_0 \qquad y = y^* a_0$$

- 如果度量误差只是一个随机的报告误差，与自变量没有关系时，在通常的假设之下，OLS估计是比较合适的估计方法

4．4．7 自变量中存在度量误差

- 考虑一个满足GM1~4的假设简单的回归模型

$$y=\beta_0+\beta_1 x_1^*+\varepsilon$$

- 如果 x_1^*是不可观测的，只能通过 x_1 来度量 x_1^*
- 参数的度量误差为 $e_1=x_1-x_1^*$
- 假设度量的平均误差为0
- 另一个假设是 ε与x_1和x_1^* 无关 $E(y\mid x_1^*,x_1)=E(y\mid x_1^*)$
- OLS估计的属性与对度量误差所作的假设有关
- 假设 e_1与x_1(可观测变量)无关,$\mathrm{cov}(x_1,e_1)=0$
- 显然有 e_1与不可观测变量x_1^* 有关

- 代入原始模型得 $y=\beta_0+\beta_1x_1+(\varepsilon-\beta_1e_1)$
- 根据假设 $(\varepsilon-\beta_1e_1)$也是期望为零且与 x_1 无关
- 得到OLS估计是 β_1和β_0 的一致估计
- 因为 ε和e_1也不相关，所以有$Var(\varepsilon-\beta_1e_1)=\sigma_\varepsilon^2+\beta_1^2\sigma_{e_1}^2$
- 比较合理的假设是度量误差与不可观测的解释变量无关，即 $\mathrm{cov}(x_1^*,e_1)=0$
- 观测的度量表为真实取值和度量误差之和$x_1=x_1^*+e_1$

$$\mathrm{cov}(x_1,e_1)=E(x_1e_1)=E(x_1^*e_1)+E(e_1^2)=0+\sigma_{e_1}^2=\sigma_{e_1}^2$$

- 代入可观测度量的模型中干扰项与解释变量是相关的

$$\mathrm{cov}(x_1,\varepsilon-\beta_1e_1)=-\beta_1\mathrm{cov}(x_1,e_1)=-\beta_1\sigma_{e_1}^2$$

- 在这种称为经典误差变量的假设之下得到的OLS回归估计将是有偏的和不一致的

$$p\lim(\hat\beta_1)=\beta_1+\mathrm{cov}(x_1,\varepsilon-\beta_1e_1)/Var(x_1)$$

$$\hat\beta_1=\beta_1-\beta_1\sigma_{e_1}^2/(\sigma_{x_1^*}^2+\sigma_{e_1}^2)=\beta_1[1-\sigma_{e_1}^2/(\sigma_{x_1^*}^2+\sigma_{e_1}^2)]$$

$$=\beta_1\sigma_{x_1^*}^2/(\sigma_{x_1^*}^2+\sigma_{e_1}^2)$$

- 因此 $p\lim\hat\beta_1$总是比 β_1更偏向于零
- 称为经典自变量误差问题OLS估计的压缩偏差
- 由度量误差带来的估计偏离依赖于度量误差的方差和不可观测变量的方差的相对大小 $Var(x_1^*)/Var(x_1)$

- 模型中包含多个自变量时

$$y=\beta_0+\beta_1x_1^*+\beta_2x_2+\beta_3x_3+\varepsilon$$

第一个自变量有度量误差，ε 与 x_1^*,x_2,x_3和x_1都不相关

- 大多数情况下，可以假设 e_1 与 x_2,x_3 都不相关
- 如果假设 e_1与 x_1不相关，则可以得到对y进行回归的OLS估计是一致的
- 在CEV假设之下估计也通常是有偏的和不一致的
- 对β_1 仍然具有估计的压缩效应

$$p\lim(\hat\beta_1)=\beta_1\sigma_{r_1^*}^2/(\sigma_{r_1^*}^2+\sigma_{e_1}^2)$$

- 只有唯一度量误差变量时有k个自变量的情况也成立
- 只有在 x_1^*与 x_2,x_3都不相关时，$\hat\beta_2$和 $\hat\beta_3$才是一致的
- 不止一个自变量存在度量误差时要给出CEV假设下的结果是非常困难的
- CEV的假设也不见得就更好，真实的情况很可能大多是介于这二者之间
- 用工具变量或GMM方法在一定的假设之下，对一般的变量存在度量误差问题，可以给出参数的一致估计

4．5．数据缺失，非随机抽样和异常值的处理

- 4．5．1缺失数据问题可以有多种表现形式
- 几乎所有的回归分析软件都有检查数据缺失的功能，
- 有一些方法来处理这些观测值中可用信息的方法
- 有时样本的剔除将导致非随机抽样的问题
  - 平均受教育比较低的人的教育数据缺失比例比较高
  - IQ比较高的人比较容易有IQ记录
- 4．5．2非机即抽样　假设2不成立时，选择的自变量在一定基准之上可以不产生统计问题
- 称为基于自变量的选样方法，是外生样本选择的一种情形

- 家庭的年存款额依赖于收入、年龄、家庭成员数或其它因素，模型可以表示为

$$saving = \beta_0 + \beta_1 income + \beta_2 age + \beta_3 size + \varepsilon$$

- 数据是对35岁以上的人进行调查而得
- 用非随机抽样得到的数据来给出有关参数的无偏的和一致的估计
- 对这一部分样本仍有E(sav|inc,age,size)对任意一组给定的样本取值具有相同的期望
- 估计不再是有效的
- 如果决定是否进入抽样范围的其它因素与干扰项独立，则仍然可以得到外生的选择

91

- 如果样本的选择是基于因变量,称为基于因变量的选样，是内生样本选择的一种情形
- 例如，在成年人总体中，个人财富与一些因素之间的关系，建立的模型为

$$wealth = \beta_0 + \beta_1 educ + \beta_2 \exp er + \varepsilon$$

- 样本中只包含了财富超过10万元的人
- 显然是所关注人群的一个非随机抽样
- 产生有偏的原因是总体回归模型所具有的条件期望为E(wealth |educ, exper, age)
- 常见的一种数据收集方式是分层抽样，把总体分成互不重叠的很多个组

92

- CEO年薪例子可能会认为年薪低于某个水平的公司可能是国有企业或者是管理层主要通过持股，不是通过年薪来反映其报酬，我们可能会想当然地设定某一个水平来剔除一部分样本，例如CEO年薪的对数大于12.2
- 使用1054家公司给出估计模型为

$$\widehat{\log(CEO)} = 6.670 + 0.280 \ln asset + 0.249 earn + 0.035 \ln income - 0.797 herfindahl$$
$$(0.411)\quad (0.023)\quad (0.046)\quad 0.011)\quad (0.202)$$

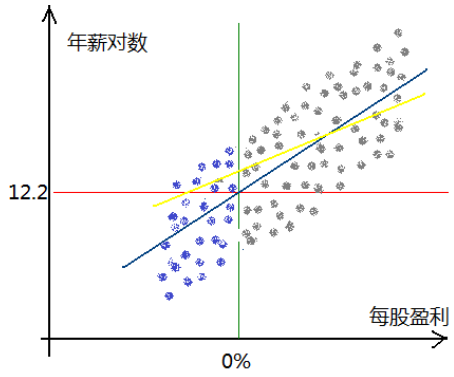
只使用CEO年薪高于12.2的954家公司给出估计模型为

$$\widehat{\log(CEO)} = 8.372 + 0.230 \ln asset + 0.322 earn + 0.008 \ln income - 0.576 herfindahl$$
$$(0.383)\quad (0.021)\quad (0.043)\quad (0.011)\quad (0.177)$$

如使用公司的盈利水平高低作为选择样本的标准
盈利为正的公司刚好955家给出估计模型为

$$\widehat{\log(CEO)} = 6.506 + 0.284 \ln asset + 0.228 earn + 0.041 \ln income - 0.862 herfindahl$$
$$(0.441)\quad (0.025)\quad (0.058)\quad (0.012)\quad (0.215)$$

93



94

- 分层抽样方法是比较明显的容易引起非随机性
- 工资和教育程度的例子中，所调查到的数据都是实际上正在工作的人群,正好是所提供的工资
- 对没有接受工作的人就无法观测到他被提供的工资
- 只能对正在工作的这部分人得到样本
- 能得到无偏估计吗？
- 决定是否工作可能会与没有观测到的公司所提供的工资水平有关
- 这一选择可能是内生的，当然也就会导致估计有偏

95

4．5．3 异常值的影响

- 样本不太大时，OLS估计的结果可能会受到个别或少数几个观测值的影响
- 最小化问题中比较大的残差有很大的权重
- 异常值可能有两个来源
- 由于出现差错而在数据中产生了异常值
- 异常也可能表现为样本量不大时，个别样本与其它样本相比差别太大
- 回归分析中是否需要剔除这些异常值是一件非常困难的事

96

- 最好把包含异常值和不包含异常值的结果都给出来
- 硕士生年薪的例子
- 通过OLS对154位同学的数据进行估计得到

$$salary = -14.069 - 0.355y2008 - 6.259y2009 - 17.573under8 - 19.026under9 + 13.271GPA$$
（14.012） （2.460） （2.541） (2.712) (2.634) (3.990)

统计方法中都是用OLS回归的残差来定义异常值，残差远远超过标准差的就认为是异常值。在前面的例子中，残差序列标准差的估计为 $\hat{\sigma} = 11.607$

残差比较高的二位同学的残差值分别是标准的2.67，-2.17倍。

在剔除了这两个异常值之后得到的估计模型为

$$salary = -22.908 - 1.00y2008 - 7.745y2009 - 18.003under8 - 19.456under9 + 15.964GPA$$
（0.641） （0.112） （0.116） (0.124) (0.121) (0.182)

97

- 有的函数形式对异常值不敏感，用对数变换来明显改善数据的变化用154位同学的数据估计得到模型

$$\ln salary = 10.861 + 0.039y2008 - 0.316y2009 - 0.698under8 - 0.757under9 + 0.455GPA$$
（0.672） （0.118） （0.122） (0.130) (0.126) (0.191)

有一位来自香港同学，其GPA只有2.0，回港工作，年薪相比其他人高，可认为是一个奇异点。对剩下的153位进行估计得

$$\ln salary = 10.015 + 0.072y2008 - 0.281y2009 - 0.714under8 - 0.768under9 + 0.682GPA$$
（0.765） （0.117） （0.121） (0.129) (0.125) (0.215)

去掉了一个异常点后，GPA的影响系数提高了50%

在前面的例子中残差比较高的三位同学的残差值分别为1.39，-1.53，-1.61，而上述香港同学的残差值只有1.13，剔除残差值最大的三位同学后151位同学数据给出的回归模型

$$\ln salary = 10.450 + 0.005y2008 - 0.343y2009 - 0.735under8 - 0.763under9 + 0.581GPA$$
（0.641） （0.112） （0.116） (0.124) (0.121) (0.182)

98

- 当且仅当$\beta_1 > 1$时RD投入强度增加与公司sales有关
- 用32家公司的数据估计得

$$\widehat{\log(rd)} = -4.378 + 1.084\log(sales) + 0.0217profmarg$$
（0.468） （0.062） （0.0128）

- 用31家公司的数据估计得

$$\widehat{\log(rd)} = -4.404 + 1.088\log(sales) + 0.0218profmarg$$
（0.511） （0.067） （0.013）

- 二者结果几乎相同，两个模型都不能拒绝 $H_0 : \beta_1 = 1$

99

- 处理观测值对模型造成重大影响的另一种方法是使用对异常值不敏感的估计方法
- 最小绝对差（least absolute deviations LAD）
- LAD方法是估计自变量对因变量的条件中位数的影响，而不是条件均值
- 当分布是对称时中位数和均值才相同
- GM假设中并没有对称性要求，在y的条件分布不对称时用OLS和LAD给出的估计可能相差很大，而这一差别可能只是因为均值和中位数本身的差别，而不是由于异常值

100

4．5．4 残差分析

- 考虑对具体某个观测值而言，实际值与估计值之间谁大谁小？这类分析称为残差分析
- 例资产的实际价格低于预测价格，价格被低估
- 通过残差分析来给出一些合理性判断，补充标准统计量的不足。

101

a). 标准化残差

- 较大的残差可以用来诊断模型
- 数据点对残差的影响
- 通过分别删除数据点来得到缺少这一点后的残差
- 模型中不包含第i个观测而得到的参数估计

$$\hat{\beta}_j(i), j = 0,1,\cdots,k$$

- 模型的残差估计为

$$\hat{\varepsilon}(i) = y_i - \hat{\beta}_0(i) - \hat{\beta}_1(i)x_{i1} - \cdots\hat{\beta}_k(i)x_{ik} \quad i = 1,\cdots,n$$

- 模型的残差方差估计为 $\hat{\sigma}^2(i) \quad i = 1,\cdots,n$

102

- 标准化而得到的残差序列为

$$\hat{\varepsilon}^*(i) = \hat{\varepsilon}(i)/\hat{\sigma}(i) \qquad i = 1, \cdots, n$$

- 样本量适当大的时，标准化残差序列服从标准正态分布
- 通过正态检验可以发现模型中是否有异常数据需要纠正或模型的设定是否适当的线索
- 残差超过2的数据点都需要给予特别注意，考虑是否异常点

103

- 硕士起薪例子

$$\ln salary = 10.234 + 0.071 y2008 - 0.344 y2009 - 0.661 under8 - 0.732 under9 + 0.337 GPA + 0.012 eco$$

（0.844） （0.120） （0.126） (0.133) (0.128) (0.208) (0.009)

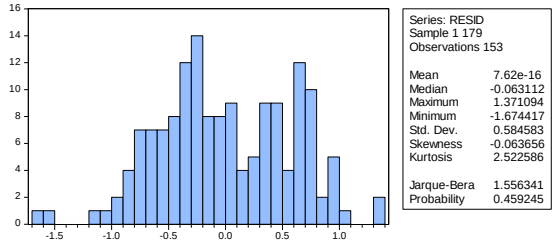


图4．2 回归模型的残差分布图

104

b)． 数据点对估计参数的影响

- 对某个参数特别感兴趣
- 考虑是否有观测值明显影响模型参数的估计结果
- 比较包含和不包含某一数据点前后模型OLS估计的差异来给出度量

$$dfbeta(i) = \frac{\hat{\beta}_j - \hat{\beta}_j(i)}{\hat{\sigma}_{\beta_j}(i)}$$

- 度量指标的绝对值大于2表明数据点是显著影响
- 数据点对估计参数的影响随样本量增加逐渐减小
- 推荐标准是使用 $\frac{2}{\sqrt{n}}$ 或 $\frac{3}{\sqrt{n}}$

105

- 硕士生起薪例子中，香港同学剔除这一样本后再给出模型的估计，使用152位同学的数据得到

$$\ln salary = 9.565 + 0.097 y2008 - 0.304 y2009 - 0.6811 under8 - 0.745 under9 + 0.57 GPA + 0.010 eco$$

（0.895） （0.120） （0.126） (0.131) (0.127) (0.234) (0.009)

模型估计系数中GPA的影响为

$$dfbeta(i) = \frac{\hat{\beta}_j - \hat{\beta}_j(i)}{\hat{\sigma}_{\beta_j}(i)} = \frac{0.57 - 0.337}{0.234} = 0.996$$

临界的标准是 $\frac{3}{\sqrt{n}} = \frac{3}{\sqrt{152}} = 0.243$

106

c). 采用对异常值稳健的估计方法

- 使用对异常值不敏感的估计方法
- LAD方法是最小化残差的绝对值和
- 没有公式解，只能采用数据计算
- LAD的目标是估计条件中位数，只有当分布是对称时中位数和均值才相同
- y的条件分布不对称时用OLS和LAD给出的估计可能相差很大
- 差别可能是因为均值和中位数本身的差别，而不是由于异常值

107

- 硕士生起薪的例子中，使用异常值稳健的LAD方法给出的估计模型为

$$\ln salary = 10.098 + 0.024 y2008 - 0.400 y2009 - 0.858 under8 - 0.830 under9 + 0.470 GPA + 0.010 eco$$

（1.676） （0.212） （0.189） (0.280) (0.272) (0.386) (0.015)

$n = 153, R^2 = 0.118, \overline{R}^2 = 0.082$

只剔除一位香港同学后得到的估计模型为

$$\ln salary = 8.313 + 0.013 y2008 - 0.456 y2009 - 0.932 under8 - 0.915 under9 + 0.949 GPA + 0.011 eco$$

（1.591） （0.198） （0.189） (0.240) (0.228) (0.337) (0.016)

$n = 153, R^2 = 0.118, \overline{R}^2 = 0.082$

把残差最大的三位同学也加入到剔除的行列，共剔除四位异常值之后的估计结果

$$\ln salary = 8.146 + 0.007 y2008 - 0.486 y2009 - 0.973 under8 - 0.941 under9 + 0.991 GPA + 0.011 eco$$

（1.586） （0.196） （0.185） (0.230) (0.216) (0.325) (0.016)

$n = 149, R^2 = 0.146, \overline{R}^2 = 0.110$

108

- 总结：
- 讨论最小二乘估计的假设条件不成立时的处理方法
- 假设6不成立，讨论使用虚拟变量的回归模型，特别是因变量为虚拟变量的线性概率模型。
- 假设5不成立，给出异方差稳健的回归模型和检验统计量，介绍异方差检验的方法，存在异方差的加权最小二乘估计方法和广义最小二乘GLS估计方法
- 假设4不成立，可能存在共线性的检验和处理方法
- 假设3不成立,函数形式设定不当检验方法和存在不可观测变量的代理变量及变量存在度量误差的估计
- 假设2不成立,数据缺失、非随机抽样和异常值处理

再见！