第五章 异方差性

引子：更为接近真实的结论是什么？

根据四川省2000年21个地市州医疗机构数与人口数资料，分析医疗机构与人口数量的关系，建立卫生医疗机构数与人口数的回归模型。对模型估计的结果如下：

$$\hat{Y}_i = -563.0548 + 5.3735 X_i$$

(291.5778)　(0.644284)

t =(-1.931062)　(8.340265)

$$R^2 = 0.785456 \quad \bar{R}^2 = 0.774146 \quad F = 69.56003$$

式中：Y表示卫生医疗机构数（个），X表示人口数量（万人）。

真的不到2000人(1860人)就需要一个医疗机构吗?

- 人口数量对应参数的标准误差较小
- t 统计量远大于临界值
- 可决系数和修正的可决系数结果较好
- F检验结果明显显著

表明该模型的估计效果不错，即可以认为人口数量每增加1万人，平均说来医疗机构将增加5.3735个。然而，这里得出的结论可靠吗?平均说来每增加1万人口真的需要增加这样多的医疗机构吗?所得结论好象并不符合实际情况。

有什么充分的理由说明这一回归结果不可靠呢？如果这一结论不可靠，更为接近真实的结论又是什么呢?

第五章 异方差性

本章将讨论四个问题：

- 异方差的实质和产生的原因
- 异方差产生的后果
- 异方差的检测方法
- 异方差的补救

第一节 异方差性的概念

一、 异方差的实质

同方差的含义

同方差性：对所有的 $i\ (i=1,2,...,n)$ 有：

$$Var(u_i|X_i)=\sigma^2$$

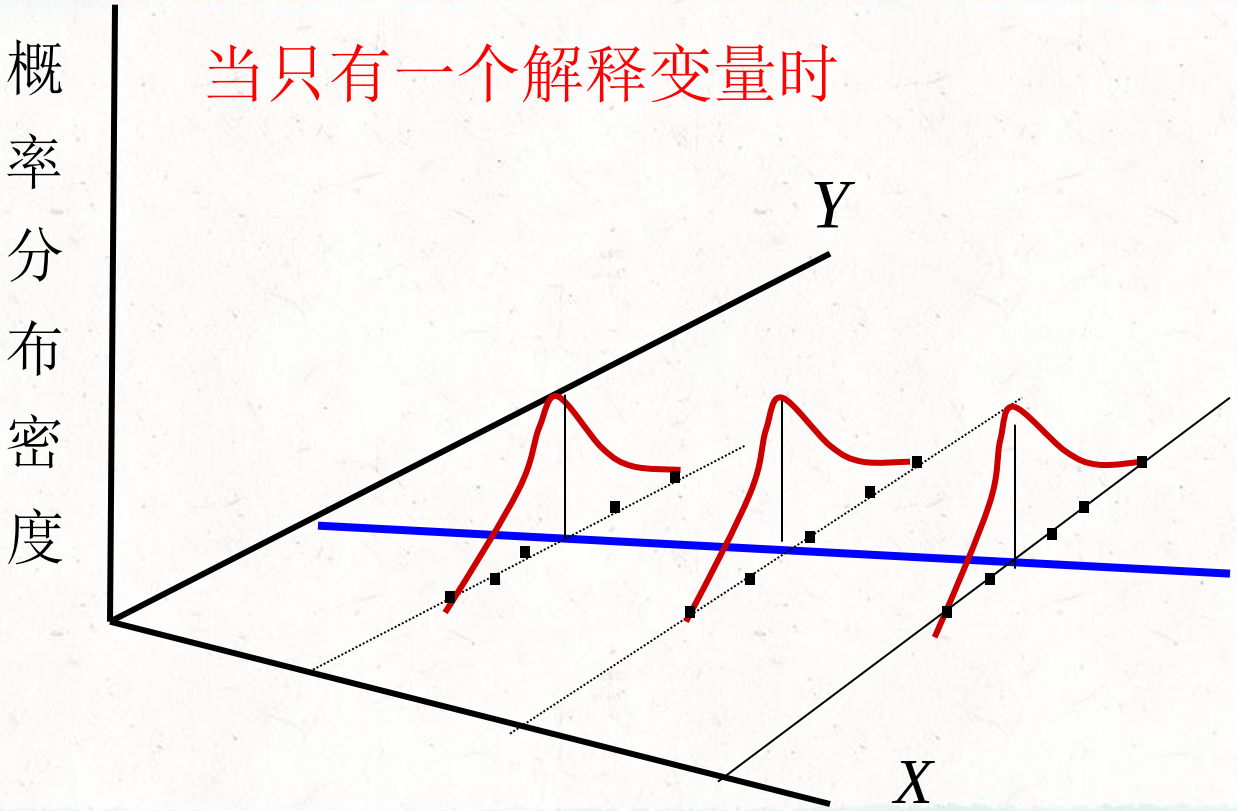
因为方差是度量被解释变量 Y 的观测值围绕条件期望

$$E(Y_i|X_2,X_3,\cdots X_k)=\beta_1+\beta_2X_{2i}+\beta_3X_{3i}+\ldots+\beta_kX_{ki}$$

的分散程度，因此同方差性指的是所有观测值围绕回归线的分散程度相同。

同方差性的图示

当只有一个解释变量时



异方差性含义

模型中的随机扰动项主要代表两方面的影响：

（1）被模型忽略的其他变量对被解释变量的影响

（2）测量误差的影响

实际上随机扰动代表的两方面因素有可能随X_i的变化而变化，使随机扰动的方差也随X_i的变化而变化，这种情况称为存在异方差性，表现为

$$Var(u_i \mid X_i) = \sigma_i^2 \qquad (i = 1,2,\cdots n)$$

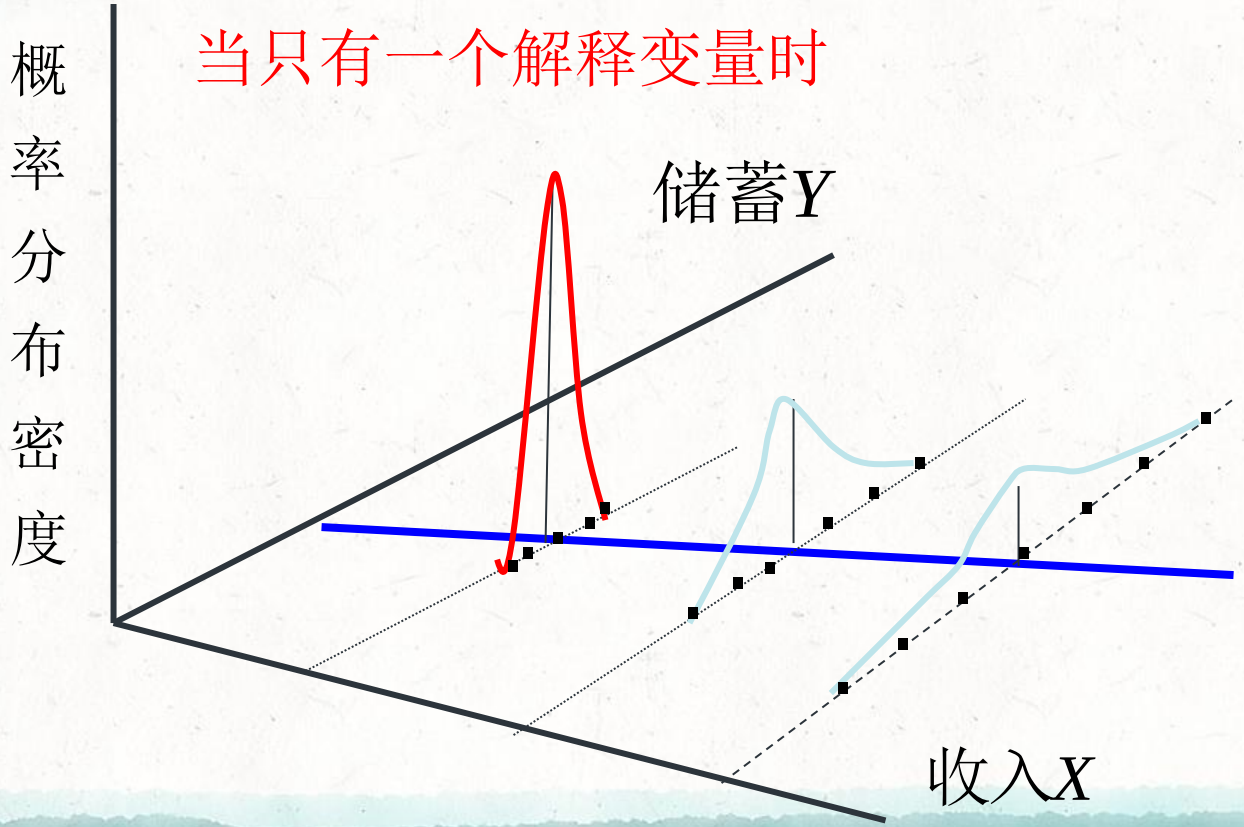
对比同方差时为 $Var(u_i \mid X_i) = \sigma^2$

异方差可看成是由于某个解释变量的变化而引起的，则

$$Var(u_i) = \sigma_i^2 = \sigma^2 f(X_i)$$

异方差性的图示

当只有一个解释变量时



二、产生异方差性的原因

- 从模型中略去的变量随列入模型的解释变量 X_i 的变化，也呈现规律性的变化，导致 u_i 的方差随 X_i 而变化
- 模型设定不恰当产生的异方差。如果一些重要变量被忽略，或把非线性模型设为线性，可能导致异方差
- 统计测量误差导致的异方差

  测量误差可能随解释变量X的增大而增大

- 截面数据中总体各单位的差异

一般说异方差性在截面数据中比在时间序列数据中更常出现

原因:同一时点不同对象的差异（如某年各省的GDP）一般大于同一对象不同时间的差异(同一个省不同年份的GDP)

注意：在经济结构发生较大变化时,时间序列也常存在异方差

第二节 异方差性的后果

存在异方差时，OLS估计仍然是无偏估计，但是

一、 OLS估计式不再具有最小方差特性

后面将证实，存在异方差时，可证明能够找到比OLS 的方差更小的估计方法。表明OLS估计式的方差不一定是最小的，即OLS估计式虽然无偏，但不一定是最佳的。

二、 解释变量的显著性检验失效

1. 参数方差的确定会有困难 ，例如一元回归时

可证明异方差时 $Var(\hat{\beta}_2^*)=\dfrac{\sum x_i^2\sigma_i^2}{(\sum x_i^2)^2}$ (证明见下页)

σ_i^2 未知，σ_i^2不再是常数，也不能再用 $(\sum e_i^2)/(n-2)=\hat{\sigma}^2$ 去估计，事实上 $Var(\hat{\beta}_2^*)$难以确定。

异方差和自相关对方差的影响

对于 $Y_i = \beta_1 + \beta_2 X_i + u_i$

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i (Y_i - \bar{Y})}{\sum x_i^2} = \frac{\sum x_i Y_i}{\sum x_i^2} - \frac{\bar{Y}\sum x_i}{\sum x_i^2} = \sum \frac{x_i}{\sum x_i^2} Y$$

$$= \sum \frac{x_i}{\sum x_i^2}(\beta_1 + \beta_2 X_i + u_i) = \beta_2 + \frac{\sum x_i u_i}{\sum x_i^2}$$

其中：$\sum \frac{x_i}{\sum x_i^2} = 0, \sum \frac{x_i X_i}{\sum x_i^2} = 1$

见教材P33(2.37)

$$Var(\hat{\beta}_2) = E(\hat{\beta}_2 - \beta_2)^2 = E[(\beta_2 + \frac{\sum x_i u_i}{\sum x_i^2}) - \beta_2]^2 = E[\frac{\sum x_i u_i}{\sum x_i^2}]^2$$

$$= E[\frac{\sum (x_i u_i)^2 + 2\sum_{i \neq j} x_i u_i x_j u_j}{(\sum x_i^2)^2}]$$

$$= \frac{\sum x_i^2 E(u_i^2) + 2\sum_{i \neq j} x_i x_j E(u_i u_j)}{(\sum x_i^2)^2}$$

回忆：

$(a+b)^2 = a^2 + 2ab + b^2$

$(\sum a_i)^2 = \sum a_i^2 + 2\sum_{i \neq j} a_i a_j$

在异方差且自相关时

$E(u_i^2)=\sigma_i^2,\qquad E(u_iu_j)\neq 0$

$$Var(\hat{\beta}_2)=\frac{\sum x_i^2E(u_i^2)+2\sum_{i\neq j}x_ix_jE(u_iu_j)}{(\sum x_i^2)^2}$$

在同方差且无自相关时

$E(u_i^2)=\sigma^2,\qquad E(u_iu_j)=0$

$$Var(\hat{\beta}_2)=\frac{\sum x_i^2E(u_i^2)}{(\sum x_i^2)^2}=\frac{\sigma^2}{\sum x_i^2}$$

在异方差但无自相关时

$E(u_i^2)=\sigma_i^2,\qquad E(u_iu_j)=0$

$$Var(\hat{\beta}_2)=\frac{\sum x_i^2E(u_i^2)}{(\sum x_i^2)^2}=\frac{\sum x_i^2\sigma_i^2}{(\sum x_i^2)^2}$$

在同方差但自相关时

$E(u_i^2)=\sigma^2,\qquad E(u_iu_j)\neq 0$

$$Var(\hat{\beta}_2)=\frac{\sigma^2}{\sum x_i^2}+\frac{2\sum_{i\neq j}x_ix_jE(u_iu_j)}{(\sum x_i^2)^2}$$

在异方差但无自相关时

$$Var(\hat{\beta}_2)=\frac{\sum x_i^2 E(u_i^2)}{(\sum x_i^2)^2}=\frac{\sum x_i^2\sigma_i^2}{(\sum x_i^2)^2}$$

设存在异方差时的参数为β_2^*，估计式为 $\hat{\beta}_2^*$

例如为 $\sigma_i^2=\sigma^2 f(X_i)=\sigma^2 X_i^2$

存在异方差时，用OLS估计的 $\hat{\beta}_2^*$ 的方差为:

$$Var(\hat{\beta}_2^*)=\frac{\sum x_i^2\sigma^2 X_i^2}{(\sum x_i^2)^2}=\frac{\sigma^2}{\sum x_i^2}\cdot\frac{\sum x_i^2 X_i^2}{\sum x_i^2}=Var(\hat{\beta}_2)\cdot\frac{\sum x_i^2 X_i^2}{\sum x_i^2}$$

因为 $(\sum x_i^2 X_i^2/\sum x_i^2)\neq 1$ 所以有 $Var(\hat{\beta}_2^*)\neq Var(\hat{\beta}_2)$

注意:$Var(\hat{\beta}_2)$是不存在异方差时$\hat{\beta}_2$的方差,$Var(\hat{\beta}_2)$具有效性(最小方差性),$\hat{\beta}_2^*$则不具有有效性。

2.如果仍然用不存在异方差性时的OLS方式估计其方差，即用

$$Var(\hat{\beta}_2) = \sigma^2 / \sum x_i^2$$

所估计的方差，可能会低估或高估存在异方差时的的真实方差。

后果：低估或高估$Var(\hat{\beta}_2)$，也就会高估或低估 t 统计量，用t检验对参数统计显著性的检验失去意义。

3. 预测精度降低，区间预测面临困难

尽管参数的OLS估计量仍然无偏，并且基于此的预测也是无偏的，但是

- 由于异方差的存在，$\hat{\beta}_k$ 的方差不再是最小的，Y的预测精度下降。
- 由于 σ_i^2 难以确定，Y的方差也难以确定，Y置信区间的确定事实上会出现困难。
- 在 $\hat{\sigma}^2 = \sum e_i^2 / n-k$ 是 σ^2 无偏估计的证明中用到了 u_i 的同方差性假定，由于存在异方差性，使得 $\hat{\sigma}^2 = \sum e_i^2 / n-k$ 是有偏的，在此基础上的区间估计和区间预测都将不可靠。

第三节 异方差性的检验

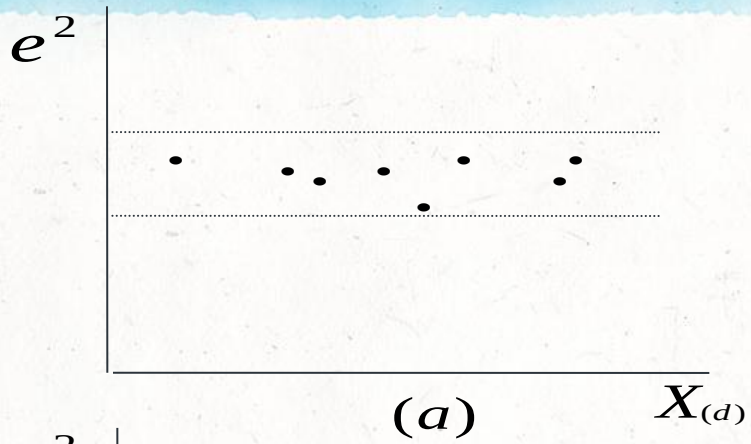
一. 图形分析法

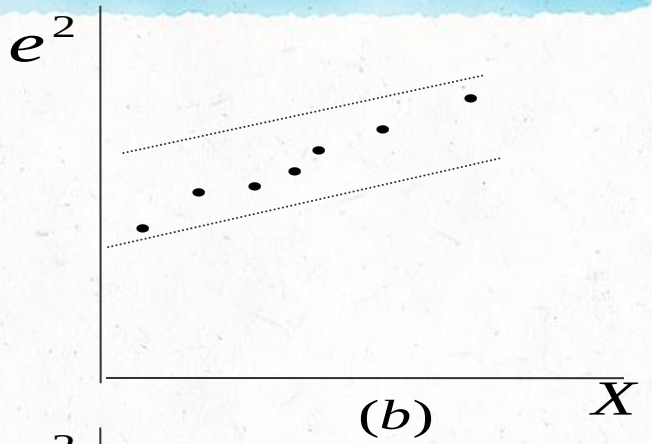
基本思想：

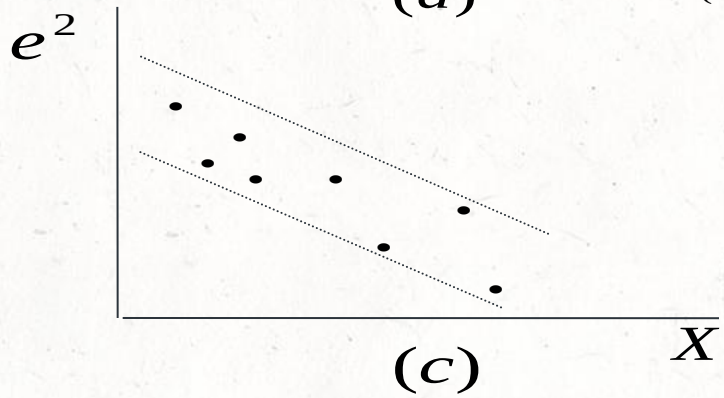
异方差性的表现是 u_i 的方差随某个解释变量的变化而变化，或Y的分散程度随X的变化而变化。因此可利用 u_i 的代表 e_i 与某解释变量的散布图，观察是否存在异方差及其异方差的形式。

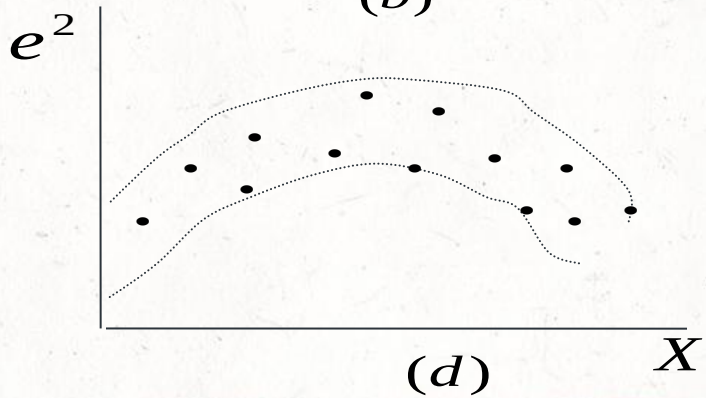
具体方法：

- 假定不存在异方差，进行回归，并计算剩余平方 e_i^2 ,描绘 e_i^2 与 X_i 的散点图，作出近似判断。
- 分析Y与X的相关图形，也可以初略地看到Y的离散程度与X之间是否有相关关系。


e^2
$X_{(d)}$
(a)

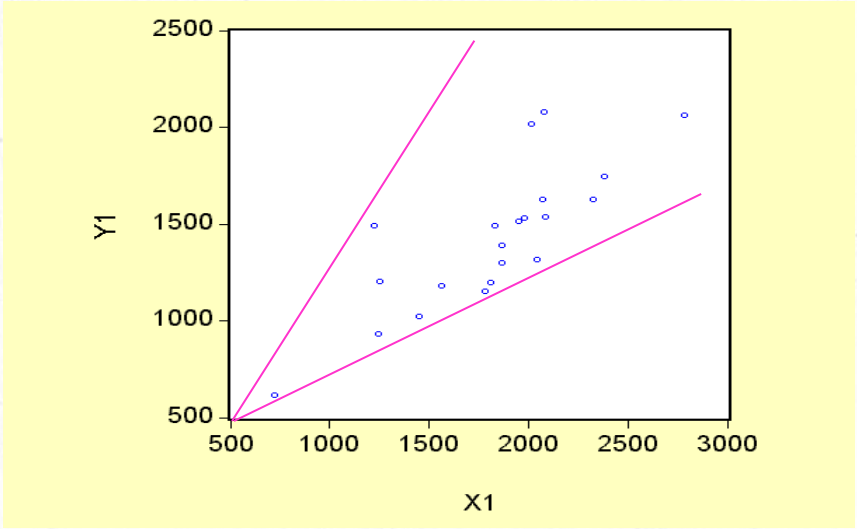

e^2
X
(b)


e^2
X
(c)


e^2
X
(d)

Y与X之间图形举例:

用1998年四川省各地市州农村居民家庭消费支出与家庭纯收入的数据，绘制出消费支出对纯收入的散点图，其中用 Y_1 表示农村家庭消费支出，X_1 表示家庭纯收入。



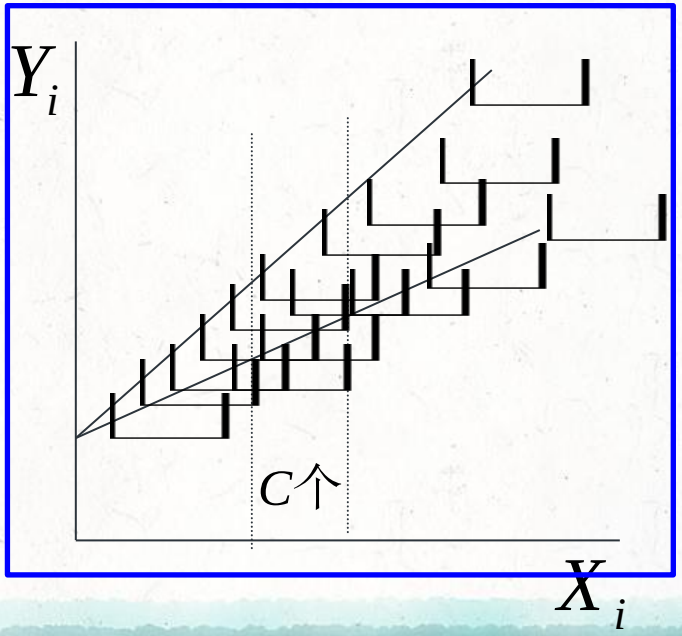
二、 Goldfeld-Quandt 检验（GQ检验）

作用：检验递增性(或递减性)异方差。

基本思想：

- 将观测值按 X_i 的大小顺序排列
- 去掉中间位置的一部分观测值，从而把观测值分为前后两部分

（目的是提高分辨性）



- 将前后两部分分别作回归，分别计算出各部分剩余 e_i ，
- 比较前后两个回归的剩余平方和 $\sum e_i^2$ ：

  如果两个 $\sum e_i^2$ 之比接近于1，为同方差；

  如果两个 $\sum e_i^2$ 之比不同于1，为异方差

前提条件：

- 样本容量较大
- u_i 服从正态分布，并除异方差外满足其他基本假定

具体步骤：

●排序:将观测值按解释变量X大小顺序排列

●数据分组:去掉中间的C个（约1/4）观测值，分别进行前后两部分 $(n-c)/2$ 个观测 值的回归

●提出假设:分别进行前后两部分回归的基础上，提出检验假设：

$H_o: u_i$ 是同方差（前后两部分方差无显著差异），

即 $H_0: \sigma_i^2 = \sigma^2$

$H_1: u_i$ 是异方差（方差随X递增或递减）

如为递增 $H_1: \sigma_1^2 \le \sigma_2^2 \le \cdots \le \sigma_n^2$

如为递减 $H_1: \sigma_1^2 \ge \sigma_2^2 \ge \cdots \ge \sigma_n^2$

●构造F统计量：

1. 若方差随X递增

统计量 $$F=\frac{\sum e_{2i}^2\Big/(\frac{n-c}{2}-k)}{\sum e_{1i}^2\Big/(\frac{n-c}{2}-k)}$$

F服从第一、二自由度均为 $[(n-c)/2]-k$ 的F 分布。

●判断：

查表得F临界值 $F_\alpha[(n-c)/2-k,(n-c)/2-k]$

◆若 $F\geq F_\alpha$（临界值），说明后部分比前部分显著大，就拒绝H_0 (同方差)，即接受存在异方差性

◆若 $F<F_\alpha$（临界值），说明后部分比前部分不显著大，就不拒绝H_0 ，认为是同方差性

2. 如果方差随X递减

统计量

$$F=\frac{\sum e_{1i}^2\Big/(\frac{n-c}{2}-k)}{\sum e_{2i}^2\Big/(\frac{n-c}{2}-k)}$$

F服从第一、二自由度均为 $[(n-c)/2]-k$ 的F 分布。

判断:查表得F临界值 $F_\alpha[(n-c)/2-k,(n-c)/2-k]$

◆若 $F\geq F_\alpha$（临界值），说明前部分比后部分显著大，就拒绝 H_0(同方差)，即接受存在异方差性

◆若 $F<F_\alpha$（临界值），说明前部分比后部分不显著大，就接受H_0，认为是同方差性

Goldfeld-Quandt 检验的特点

- 要求大样本
- 异方差的表现既可为递增型，也可为递减型
- 检验结果与选择数据删除的个数c的大小有关
- 只能判断异方差是否存在，在多个解释变量的情况下，对是哪一个变量引起异方差的判断存在局限（注意这两个字的理解）。

基本思想可以推广：比较不同样本的残差平方和是否有明显差异？

三、 White检验

基本思想：

如果存在异方差，其方差σ_t^2 与某解释变量有关系。在不知道关于异方差的任何先验信息时，在大样本的情况下，将OLS估计后的残差平方对解释变量的各种形式(如常数、解释变量、解释变量的平方及其交叉乘积等)构成一个辅助回归，利用辅助回归建立相应的检验统计量来判断异方差性。

例如两个解释变量的模型中 $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$

设 σ_t^2 与 X_2和X_3的关系为如下辅助回归:

$$\sigma_t^2 = \alpha_1 + \alpha_2 X_{2t} + \alpha_3 X_{3t} + \alpha_4 X_{2t}^2 + \alpha_5 X_{3t}^2 + \alpha_6 X_{2t} X_{3t} + v_t$$

其中 v_t 为随机误差项。

但一般σ_t^2未知，可用原模型回归剩余 e_t^2作为σ_t^2的估计值，进行以上辅助回归。在大样本情况下寻求能确定分布的统计量，判断 σ_t^2的变化是否与解释变量有关

(当有K个解释变量时，可作类似的含两两交互的辅助回归)

检验的基本步骤：

1. 估计原模型并计算 e_t^2

用OLS法估计原模型，计算残差$e_t = Y_t - \hat{Y}_t$，并求残差的平方e_t^2 。生成新变量 e_t^2

2. 求辅助函数

用残差平方 e_t^2 作为异方差 σ_t^2 的估计，并建立对 $X_{2t}, X_{3t}, X_{2t}^2, X_{3t}^2, X_{2t}X_{3t}$ 的辅助回归，即

$$\hat{e}_t^2 = \hat{\alpha}_1 + \hat{\alpha}_2 X_{2t} + \hat{\alpha}_3 X_{3t} + \hat{\alpha}_4 X_{2t}^2 + \hat{\alpha}_5 X_{3t}^2 + \hat{\alpha}_6 X_{2t} X_{3t}$$

并计算辅助回归的R^2

3. 提出假设

$$H_0: \alpha_2 = ... = \alpha_6 = 0, \qquad H_1: \alpha_j (j=2,,3,...,6)\ 不全为零$$

4. 计算统计量 nR^2

n 为样本容量，R^2 为辅助回归可决系数

在大样本情况下可以证明，在零假设成立下，nR^2 服从自由度为5的 χ^2 分布，即 $nR^2 \sim \chi^2(5)$

5.检验

给定显著性水平 α，查 χ^2 分布表得临界值 $\chi^2_\alpha(5)$，

如果 $nR^2 \geq \chi^2_\alpha(5)$，$H_0$ 不合理，则拒绝原假设H_0，

即认为模型中随机误差存在异方差 。

如果 $nR^2 < \chi^2_\alpha(5)$ 则不拒绝 H_0 ，即认为模型中

随机误差是同方差。

White检验的特点

- 要求为大样本
- 不仅能够检验异方差的存在性，同时在多变量的 情况下，还能判断出是哪一些变量引起的异方差。
- 缺陷：对于多个解释变量的模型，运用该检验时辅助回归会丧失较多的自由度（改进思路见P119-120)。

四、 ARCH检验(补充)

什么是ARCH(autoregressive conditionanl heteroskedasticity 自回归条件异方差）过程？ 即 $\sigma_t^2 = \alpha_0 + \alpha_1 \sigma_{t-1}^2 + \cdots + \alpha_p \sigma_{t-p}^2 + v_t$

p为ARCH过程的阶数，并且 $\alpha_0 > 0, \alpha_i \geq 0, i = 1, 2, \cdots, p$

v_t 为随机误差。

基本思想：

- 时间序列数据中可认为存在的异方差性为ARCH过程。
- 可以检验ARCH过程是否成立去判断是否存在异方差。
- 因各个σ_t^2 均未知，用对原模型OLS估计的剩余项平方e_t^2 去近似估计。

ARCH 检验的基本步骤

1. 估计参数并计算e_t

用OLS法估计原模型参数，求出残差e_t，并计算残差平方序列$e_t^2, e_{t-1}^2, \ldots, e_{t-p}^2$，以分别作为对$\sigma_t^2, \sigma_{t-1}^2, \ldots, \sigma_{t-p}^2$的估计。

2. 作ARCH过程辅助回归

$$\hat{e}_t^2 = \hat{\alpha}_0 + \hat{\alpha}_1 e_{t-1}^2 + \ldots + \hat{\alpha}_p e_{t-p}^2$$

计算辅助回归的可决系数R^2

3. 提出原假设 $H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_p = 0$；　　$H_1: \alpha_j$不全为零

4. 检验

计算统计量辅助回归的可决系数 R^2 与 $n-p$ 的乘积 $(n-p)R^2$ ，在 H_0 成立时，可以证明基于大样本，统计量 $(n-p)R^2$ 渐近服从 χ^2 分布，即 $(n-p)R^2 \sim \chi_p^2$

给定显著性水平 α ，查 χ^2 分布表得临界值 $\chi_\alpha^2(p)$

- 如果 $(n-p)R^2 \geq \chi_\alpha^2(p)$，则拒绝 $H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_p = 0$，说明ARCH过程成立，表明模型存在异方差。
- 如果 $(n-p)R^2 < \chi_\alpha^2(p)$ ，则不拒绝 $H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_p = 0$，说明ARCH过程不成立，表明模型不存在异方差。

ARCH检验的特点

- 适于时间序列数据
- 变量的取值为大样本，
- 只能判断模型中是否存在异方差，而不能诊断出哪一个变量引起的异方差。

五、 Glejser检验

检验的基本思想

用OLS法将被解释变量Y对解释变量回归得残差e_i，取e_i^2或e_i的绝对值$|e_i|$，然后将e_i^2或$|e_i|$对某个解释变量X_i的各种函数形式回归，(用各种函数形式去试，寻找最佳的函数形式)，例如$e_i^2 = f(X_i)+v_i$ 或 $|e_i| = f(X_i)+v_i$

分别检验此回归系数的显著性,若回归系数显著，就说明存在异方差，若回归系数都不显著，就认为是同方差

注意：用e_i的绝对值$|e_i|$而不用实际值，原因是：OLS中$\sum e_i X_i = 0$，无法进行$e_i = f(X_i)$的回归。

检验的步骤

1. 根据样本数据建立回归模型，作OLS回归，并计算残差序列 $e_i = Y_i - \hat{Y}_i$

2. 计算 e_i^2 或 $|e_i|$

3. 用 e_i^2 或 $|e_i|$ 对 X 的各种函数进行回归，用各种函数形式去试验，寻找最佳的函数形式。

例如 $f(X_i) = X_i$ $f(X_i) = X_i^2$ $f(X_i) = (\alpha_1 + \alpha_2 X_i)^2$

4. 判断：

用回归所得到的 β 、t 、F 等信息判断，若参数 β 显著不为零，即认为存在异方差性。

检验的特点

不仅能对异方差的存在进行判断，而且还能对异方差随某个解释变量变化的函数形式 进行诊断。

该检验要求变量的观测值为大样本。

问题：

- e_i^2或$|e_i|$与 X_i回归的函数形式事实上不可能一一 找完。
- $|e_i| = f(X_i)$回归的误差项 v_i 本身可能出现均值不为0、自相关、异方差
- 一般只可用于大样本的情况（软件使用nR2统计量）

第四节 异方差的修正

一、对原模型加以变换

基本思想： 原模型为 $Y_i = \beta_1 + \beta_2 X_i + u_i$

- u_i的异方差性与 X_i的变化有关，假定 $\sigma_i^2 = K^2 f(X_i)$

其中的 K^2为常数

- 如果以 $\sqrt{f(X_i)}$ 除原模型两边，将模型变换为

$$\frac{Y_i}{\sqrt{f(X_i)}} = \frac{\beta_1}{\sqrt{f(X_i)}} + \beta_2 \frac{X_i}{\sqrt{f(X_i)}} + \frac{u_i}{\sqrt{f(X_i)}}$$

变换后的模型的扰动项 $\frac{u_i}{\sqrt{f(X_i)}} = v_i$ 是同方差的，因为

$$Var(v_i) = Var(\frac{u_i}{\sqrt{f(X_i)}}) = \frac{1}{f(X_i)} Var(u_i) = \frac{k^2 f(X_i)}{f(X_i)} = k^2$$

具体作法：对 $f(X_i)$ 的函数形式可作出各种假定，

例如：

函数形式	$\mathrm{var}(u_i)$	v_i	$\mathrm{var}(v_i)$
$f(X_i)=X_i$	k^2X_i	$u_i/\sqrt{X_i}$	k^2
$f(X_i)=X_i^2$	$k^2X_i^2$	u_i/X_i	k^2
$f(X_i)=(\alpha_1+\alpha_2X_i)^2$	$k^2(a_0+a_1X_i)^2$	$u_i/(a_0+a_1X_i)$	k^2

注意：● $f(X_i)$ 的函数形式可参考图形分析法或**Glejser**法去确定

●模型变换可能引起变量出现“虚构的”的相关关系

●对原模型变换后的拟合优度可能变小，这是对观测值加权的结果

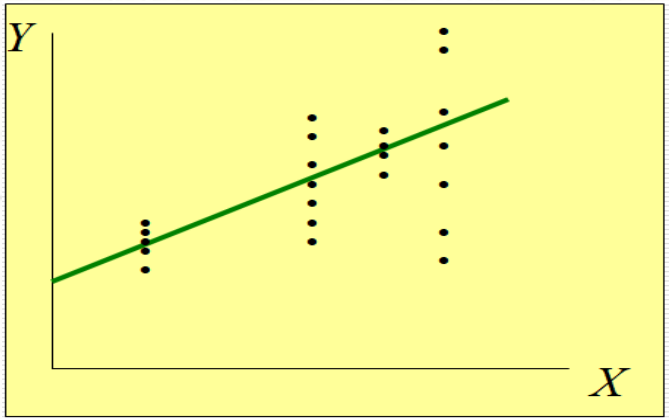
二、 加权最小二乘法(WLS)

基本思想：

●用OLS法估计参数时是使 $\min: \sum e_i^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$

这里不论 e_i^2 的大小是同等对待的（因是同方差）。

●在异方差的情况下，方差越小，其样本值偏离条件均值的程度越小，其观测值越应受到重视。即方差越小，在确定回归线时的作用应当越大，反之方差越大，其观测值所起的作用应当越小。



所以，在方差 σ_i^2已知的情况下，可以用 $1/\sigma_i^2$作为权数，使得

$$\min : \sum \frac{e_i^2}{\sigma_i^2} = \sum \frac{1}{\sigma_i^2}(Y_i - \beta_1^* - \beta_2^* X_i)^2$$

若令 $1/\sigma_i^2 = w_i$ 即

$$\min : \sum w_i e_i^2 = \sum w_i (Y_i - \beta_1^* - \beta_2^* X_i)^2$$

按这样的原则估计的 β_1^*和β_2^* 称为加权最小二乘估计式(WLS)

具体作法：如果σ_i^2已知，令 $1/\sigma_i^2 = w_i$可以证明WLS估计为

$$\beta_2^* = \frac{\sum w_i y_i^* x_i^*}{\sum w_i x_i^{*2}} \qquad \beta_1^* = \overline{Y^*} - \beta_2^* \overline{X^*}$$

其中：$\overline{Y^*}$、$\overline{X^*}$为加权平均数

$$\overline{Y^*} = \frac{\sum w_i Y_i}{\sum w_i} \qquad \overline{X^*} = \frac{\sum w_i X_i}{\sum w_i}$$

y_i^*，x_i^*为与加权平均数的离差

$$y_i^* = Y_i - \overline{Y^*} \qquad x_i^* = X_i - \overline{X^*}$$

可以证明，加权最小二乘估计可以消除或减轻异方差的影响。

模型变换与加权最小二乘法的关系

例如原模型 $Y_i = \beta_1 + \beta_2 X_i + u_i$

如果 u_i 为异方差，假定其方差为 $Var(u_i) = \sigma_i^2 = k^2 X_i^2$

1. 模型变换

此时 $f(X_i) = X_i^2$ $\sqrt{f(X_i)} = X_i$

原模型变换为 $\frac{Y_i}{X_i} = \frac{\beta_1}{X_i} + \beta_2 + \frac{u_i}{X_i}$

其中随机项 u_i/X_i 是同方差的，用OLS法估计参数，变换后的剩余平方和为 $\sum e_i^2 = \sum (\frac{Y_i}{X_i} - \frac{\hat{\beta}_1}{X_i} - \hat{\beta}_2)^2 = \sum \frac{1}{X_i^2} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$

2. 加权最小二乘法

$Var(u_i)=\sigma_i^2=k^2X_i^2$，用 u_i 方差的倒数 $\frac{1}{\sigma_i^2}=\frac{1}{k^2X_i^2}$ 作为权数

其加权最小二乘回归的剩余平方和为 $\sum(e^{*2}_i/\sigma_i^2)$

$$\sum(\frac{e^{*2}_i}{\sigma_i^2})=\sum\frac{1}{\sigma_i^2}(Y_i-\beta_1^*-\beta_2^*X_i)^2=\sum\frac{1}{k^2X_i^2}(Y_i-\beta_1^*-\beta_2^*X_i)^2$$

对比上页模型变换后的剩余平方和

$$\sum e_i^2=\sum(\frac{Y_i}{X_i}-\frac{\hat\beta_1}{X_i}-\hat\beta_2)^2=\sum\frac{1}{X_i^2}(Y_i-\hat\beta_1-\hat\beta_2X_i)^2$$

二者的剩余平方和只相差常数 k^2，能使其中一个最小的参数估计结果，必能使另一个也最小。用模型变换后的OLS估计的参数实际与应用加权最小二乘法估计的参数是一致的。

(这也间接证明了加权最小二乘法可消除异方差)

三、模型的对数变换

基本思想：

- 对数变换可使所测量变量的尺度缩小，从而缩小原变量差异的倍数,如log(10)=1; log(100)=2;LOG(1000)=3
- 对数变换后模型的剩余 e_i 表示一种相对误差，一般相对误差比绝对误差有较小的数值差异

具体作法：

原模型为： $Y_i = \beta_1 + \beta_2 X_i + u_i$

变换为 $\ln Y_i = \beta_1 + \beta_2 \ln X_i + v_i$

注意：变换后模型中参数的意义发生了变化，这时 β_2 是 Y_i 关于 X_i 的弹性，即Y相对于X的百分比变化，这与原模型中不同。

第五节　案例分析

呼应引子:医疗机构与人口数量的关系

1. 问题的提出和模型设定

为了给制定医疗机构的规划提供依据，分析比较医疗机构与人口数量的关系，建立卫生医疗机构数与人口数的回归模型。(见引子)

假定医疗机构数与人口数之间满足线性约束，则理论模型设定为：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

其中Y_i表示卫生医疗机构数，X_i表示人口数。

四川省2000年各地区医疗机构数与人口数

地区	人口数（万人）X	医疗机构数（个）Y	地区	人口数（万人）X	医疗机构数（个）Y
成都	1013.3	6304	眉山	339.9	827
自贡	315	911	宜宾	508.5	1530
攀枝花	103	934	广安	438.6	1589
泸州	463.7	1297	达州	620.1	2403
德阳	379.3	1085	雅安	149.8	866
绵阳	518.4	1616	巴中	346.7	1223
广元	302.6	1021	资阳	488.4	1361
遂宁	371	1375	阿坝	82.9	536
内江	419.9	1212	甘孜	88.9	594
乐山	345.9	1132	凉山	402.4	1471
南充	709.2	4064			

2. 参数估计

Dependent Variable: Y
Method: Least Squares
Date: 09/16/19 Time: 16:42
Sample: 1 21
Included observations: 21

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-562.9074	291.5642	-1.930646	0.0686
X	5.372828	0.644239	8.339811	0.0000

R-squared	0.785438	Mean dependent var	1588.143
Adjusted R-squared	0.774145	S.D. dependent var	1310.975
S.E. of regression	623.0301	Akaike info criterion	15.79746
Sum squared resid	7375164.	Schwarz criterion	15.89694
Log likelihood	-163.8733	Hannan-Quinn criter.	15.81905
F-statistic	69.55245	Durbin-Watson stat	0.430291
Prob(F-statistic)	0.000000		

估计结果为

$$\hat{Y}_i = -562.9074 + 5.3728 X_i$$
$$(-1.9306) \qquad (8.3398)$$
$$R^2 = 0.7854, \quad F = 69.55$$

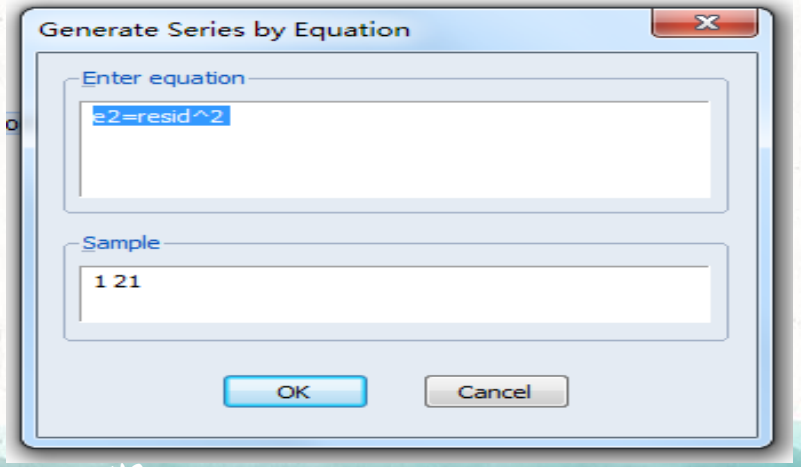
3. 检验模型的异方差

（1）图形法

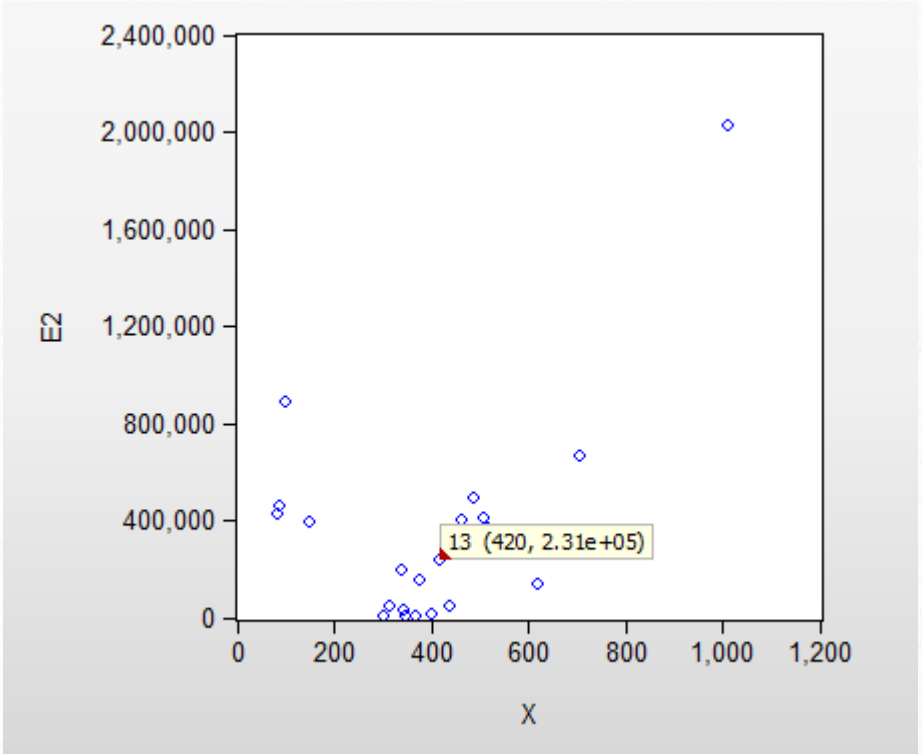
1）生成残差平方序列。在得到估计结果后，立即用生成命令建立e^2序列，记为e2。生成过程：按路径Quick /Generate Series ，进入Generate Series by Equation对话框，在Generate Series by Equation对话框中键入

“e2=resid^2”，

则生成序列e^2 。



2）绘制 e_t^2 对 X_t 的散点图。选择变量名X与e2（注意选择变量的顺序，先选的变量将在图形中表示横轴，后选的变量表示纵轴），进入数据列表，再按路径view/graph/scatter，可得散点图，见图。



判断:

由图可以看出，残差平方 e_i^2 对解释变量X的散点图主要分布在图形中的下三角部分，大致看出残差平方 e_i^2 随 X_i 的变动呈增大的趋势，因此，模型很可能存在异方差。但是否确实存在异方差还应通过更进一步的检验。

（2）Goldfeld-Quanadt检验

EViews软件操作

1）对变量取值排序（按递增或递减）。在Procs菜单里选Sort Current Page命令，出现排序对话框Sort Workfile Series，键入X，以递增型排序，选"Ascenging"，如果以递减型排序，则应选"Descending"，点ok。本例选递增型排序，这时变量Y与X将以X按递增型排序。

2）构造子样本区间，建立回归模型。在本例中，样本容量n=21，删除中间1/4的观测值，即大约5个观测值，余下部分平分得两个样本区间：1—8和14—21，它们的样本个数均是8个，即 $n_1 = n_2 = 8$。

在Sample菜单里，将区间定义为1—8，然后用OLS方法求得如右结果

表A

Dependent Variable: Y
Method: Least Squares
Date: 09/16/19 Time: 16:56
Sample: 1 8
Included observations: 8

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	598.2525	119.2922	5.015018	0.0024
X	1.177650	0.490187	2.402452	0.0531

R-squared	0.490306	Mean dependent var	852.6250
Adjusted R-squared	0.405357	S.D. dependent var	201.5667
S.E. of regression	155.4343	Akaike info criterion	13.14264
Sum squared resid	144958.9	Schwarz criterion	13.16250
Log likelihood	-50.57056	Hannan-Quinn criter.	13.00869
F-statistic	5.771775	Durbin-Watson stat	1.656269
Prob(F-statistic)	0.053117		

在Sample菜单里,将区间定义为14—21，再用OLS方法求得如右结果

表B

Dependent Variable: Y
Method: Least Squares
Date: 09/16/19 Time: 16:58
Sample: 14 21
Included observations: 8

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-2940.426	430.7787	-6.825839	0.0005
X	9.177641	0.693419	13.23534	0.0000

R-squared	0.966883	Mean dependent var	2520.500
Adjusted R-squared	0.961363	S.D. dependent var	1781.627
S.E. of regression	350.2011	Akaike info criterion	14.76721
Sum squared resid	735844.7	Schwarz criterion	14.78707
Log likelihood	-57.06884	Hannan-Quinn criter.	14.63326
F-statistic	175.1744	Durbin-Watson stat	1.815102
Prob(F-statistic)	0.000011		

3）求F统计量值。基于表中残差平方和的数据，即Sum squared resid的值。由表A计算得到的残差平方和为$\sum e_{1i}^2 = 144958.9$，由表B计算得到的残差平方和为$\sum e_{21}^2 = 735844.7$，根据Goldfeld-Quanadt检验，F统计量为

$$F = \frac{\sum e_{2i}^2}{\sum e_{1i}^2} = \frac{735844.7}{144958.9} = 5.0762$$

4）判断

在$\alpha = 0.05$下，分子、分母的自由度均为6，查F分布表得临界值为 $F_{0.05}(6,6) = 4.28$

因为 $F = 5.0762 > F_{0.05}(6,6) = 4.28$

所以拒绝原假设，表明模型确实存在异方差。

(3) White检验

在OLS估计结果界面，按路径“View/ Residual Diagnostics/ Heteroskedasticity Tests…”，在弹出的specification对话框中的检验类型中“Test type”选择“White”检验。

根据White检验中辅助函数的构造，最后几项为变量的交叉乘积项，因为本例为一元函数，故EViews7及以前版本应选择无交叉乘积项复选框。但是在EViews8中，把常数项也视为一个变量，所以常数项和解释变量X的交叉项实为X的一次项，因此应该选中“Include White cross terms”复选框，则辅助函数为

$$\sigma_i^2 = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + v_i$$

经估计出现White检验结果，见下页表。

从表中可以看出，

$nR^2 = 18.0748$

由White检验知，

在 $\alpha = 0.05$ 下，

查 χ^2 分布表得临界值，

$\chi^2_{0.05}(2) = 5.9915$

或分析P值

Heteroskedasticity Test: White

F-statistic	55.61118	Prob. F(2,18)	0.0000
Obs*R-squared	18.07481	Prob. Chi-Square(2)	0.0001
Scaled explained SS	11.78770	Prob. Chi-Square(2)	0.0028

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 09/17/19 Time: 09:25
Sample: 1 21
Included observations: 21

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	823375.5	130273.4	6.320365	0.0000
X^2	4.742387	0.532352	8.908366	0.0000
X	-3605.578	553.5894	-6.513091	0.0000

R-squared	0.860705	Mean dependent var	351198.3
Adjusted R-squared	0.845228	S.D. dependent var	454261.0
S.E. of regression	178711.1	Akaike info criterion	27.15649
Sum squared resid	5.75E+11	Schwarz criterion	27.30571
Log likelihood	-282.1432	Hannan-Quinn criter.	27.18888
F-statistic	55.61118	Durbin-Watson stat	1.687985
Prob(F-statistic)	0.000000		

对于 $H_0: \alpha_1 = \alpha_2 = 0,$ $H_1: \alpha_j (j=1,2)$ 不全为零

因为 $nR^2 = 18.0748 > \chi^2_{0.05}(2) = 5.9915$

所以拒绝原假设，不拒绝备择假设，表明模型存在异方差。

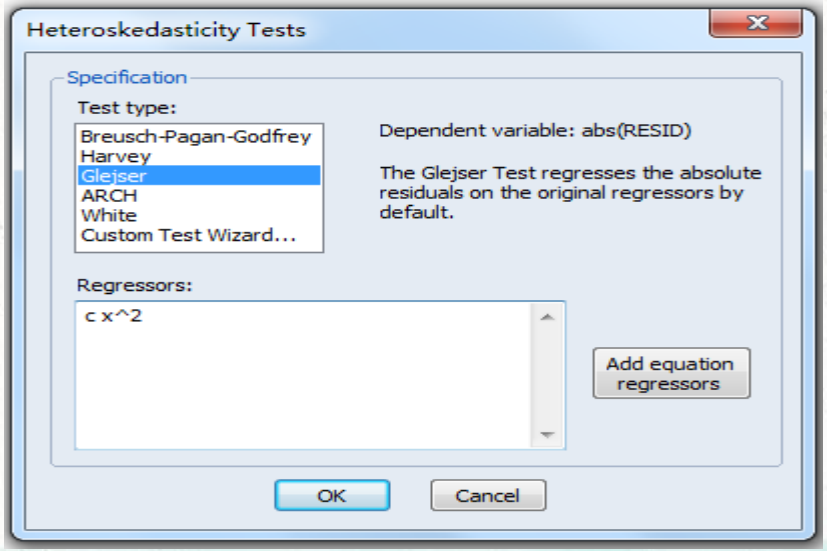
(3) Glejser检验

同样在OLS估计结果界面，按路径“View/ Residual Diagnostics/ Heteroskedasticity Tests…”，在弹出的specification对话框中的检验类型中“Test type”选择“Glejser”检验。

Glejser检验可以尝试不同的函数形式，如果以如下形式为例，

$$|e_i| = \beta_1 + \beta_2 X_i^2 + \upsilon_i$$

则在解释变量框“Regressors”中写入“C X^2”，选择OK，即可得检验结果。其他函数形式类似操作。



从表中可以看出，

$nR^2 = 6.938512$

由Glejser检验知，

在 $\alpha = 0.05$ 下，

查 χ^2 分布表得临界值，

$\chi^2_{0.05}(1) = 3.8416$

或分析P值

Heteroskedasticity Test: Glejser

F-statistic	9.375375	Prob. F(1,19)	0.0064
Obs*R-squared	6.938512	Prob. Chi-Square(1)	0.0084
Scaled explained SS	5.490546	Prob. Chi-Square(1)	0.0191

Test Equation:
Dependent Variable: ARESID
Method: Least Squares
Date: 09/17/19 Time: 09:43
Sample: 1 21
Included observations: 21

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	310.2365	85.79101	3.616189	0.0018
X^2	0.000875	0.000286	3.061923	0.0064

R-squared	0.330405	Mean dependent var	489.4709
Adjusted R-squared	0.295163	S.D. dependent var	342.3410
S.E. of regression	287.4108	Akaike info criterion	14.25010
Sum squared resid	1569495.	Schwarz criterion	14.34957
Log likelihood	-147.6260	Hannan-Quinn criter.	14.27168
F-statistic	9.375375	Durbin-Watson stat	0.998686
Prob(F-statistic)	0.006417		

因为 $nR^2 = 6.938512 > \chi^2_{0.05}(1) = 3.8416$

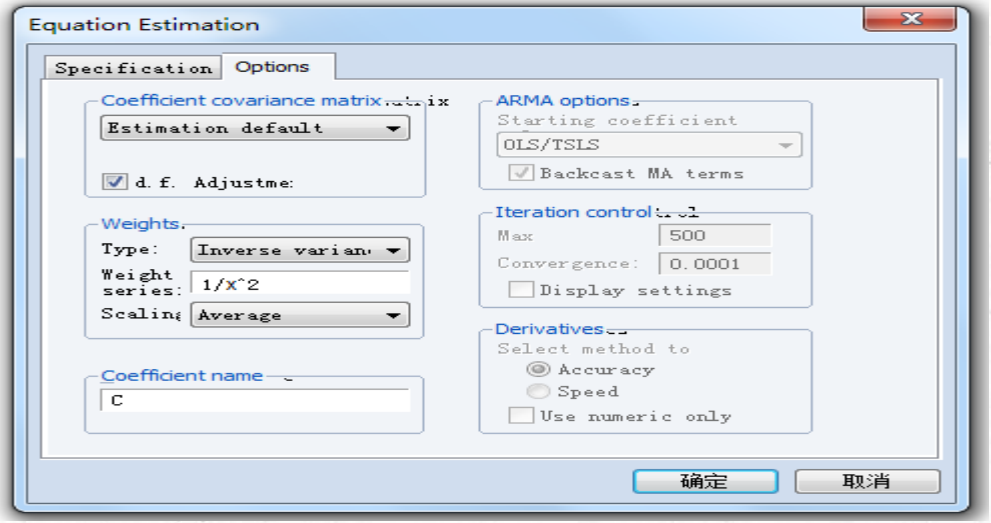
所以拒绝原假设，不拒绝备择假设，表明模型存在异方差。

4. 异方差的修正

加权最小二乘法（WLS）

分别选用权数 $w_{1t}=\frac{1}{X_t}, w_{2t}=\frac{1}{X_t^2}, w_{3t}=\frac{1}{\sqrt{X_t}}$

方法：在Quick/Estimate equation中输入“Y C X”，点option，在对话框中Weight选择加权类型“Inverse variance”、加权序列输入“1/X^2”（或者“1/X”或者“1/sqr(x)”），点击OK，即出现加权最小二乘结果。



Heteroskedasticity Test: White

F-statistic	0.980600	Prob. F(2,18)	0.3943
Obs*R-squared	2.063263	Prob. Chi-Square(2)	0.3564
Scaled explained SS	2.633894	Prob. Chi-Square(2)	0.2680

Test Equation:
Dependent Variable: WGT_RESID^2
Method: Least Squares
Date: 09/17/19 Time: 10:05
Sample: 1 21
Included observations: 21
Collinear test regressors dropped from specification

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	178180.1	91745.41	1.942114	0.0679
X*WGT^2	-1495.930	1081.901	-1.382687	0.1837
WGT^2	115040.7	82151.71	1.400345	0.1784

R-squared	0.098251	Mean dependent var	60296.47
Adjusted R-squared	-0.001944	S.D. dependent var	109116.2
S.E. of regression	109222.2	Akaike info criterion	26.17172
Sum squared resid	2.15E+11	Schwarz criterion	26.32094
Log likelihood	-271.8031	Hannan-Quinn criter.	26.20410
F-statistic	0.980600	Durbin-Watson stat	1.126059
Prob(F-statistic)	0.394251		

基于1/X^2后white检验结果

Heteroskedasticity Test: White

F-statistic	22.85406	Prob. F(2,18)	0.0000
Obs*R-squared	15.06669	Prob. Chi-Square(2)	0.0005
Scaled explained SS	21.41135	Prob. Chi-Square(2)	0.0000

基于1/X后white检验结果

Heteroskedasticity Test: White

F-statistic	55.24437	Prob. F(3,17)	0.0000
Obs*R-squared	19.04633	Prob. Chi-Square(3)	0.0003
Scaled explained SS	20.24393	Prob. Chi-Square(3)	0.0002

基于1/sqr(X)后white检验结果

注意：每尝试一种权重进行WLS后，一定要先进行异方差检验，能够通过异方差检验的WLS结果才可以进行经济含义解读。不用非要比较哪种权重最好，其实只要找到一个能消除异方差的权重即可。

Dependent Variable: Y
Method: Least Squares
Date: 09/17/19 Time: 10:16
Sample: 1 21
Included observations: 21
Weighting series: 1/X^2
Weight type: Inverse variance (average scaling)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	384.6123	87.90442	4.375346	0.0003
X	2.723571	0.433389	6.284353	0.0000

Weighted Statistics

R-squared	0.675175	Mean dependent var	847.6753
Adjusted R-squared	0.658079	S.D. dependent var	356.5126
S.E. of regression	258.1540	Akaike info criterion	14.03538
Sum squared resid	1266226.	Schwarz criterion	14.13486
Log likelihood	-145.3715	Hannan-Quinn criter.	14.05697
F-statistic	39.49310	Durbin-Watson stat	0.787378
Prob(F-statistic)	0.000005	Weighted mean dep.	808.6869

Unweighted Statistics

R-squared	0.586654	Mean dependent var	1588.143
Adjusted R-squared	0.564899	S.D. dependent var	1310.975
S.E. of regression	864.7480	Sum squared resid	14207993
Durbin-Watson stat	0.362833		

注：汇报模型整体结果时应该用“Unweighted Statistics”里相关统计量（与书上不同）。

估计结果：

$$\hat{Y}_i = 384.6123 + 2.7236 X_i$$

$$(4.3753) \quad (6.2844)$$

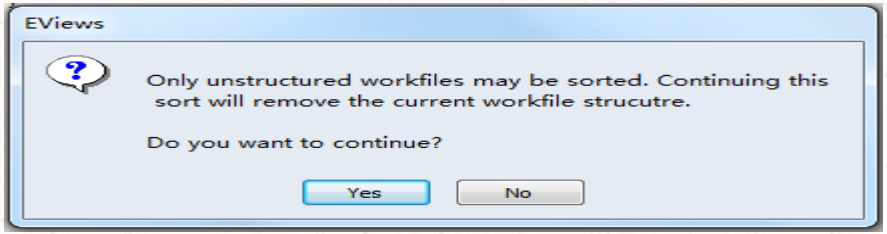
$$R^2 = 0.5867, D.W. = 0.3628$$

结论：运用加权小二乘法消除了异方差性后，参数的t检验均显著，F检验也显著，并说明人口数量每增加1万人，平均说来将增加2.7236个卫生医疗机构，而不是引子中得出的增加5.3728个医疗机构。

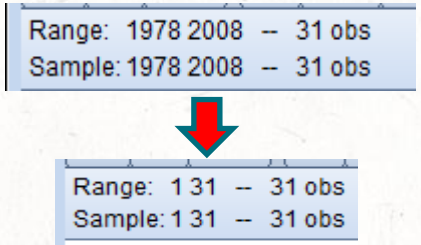
拓展：时间序列数据的情形

1. G-Q检验的不同之处

在Procs菜单里选Sort Current Page命令时，会出现以下警示框：



点击“yes”继续即可，其他操作与横截面数据类似，但排序后数据结构会变为横截面结构（如右图），所以对于时间序列数据，

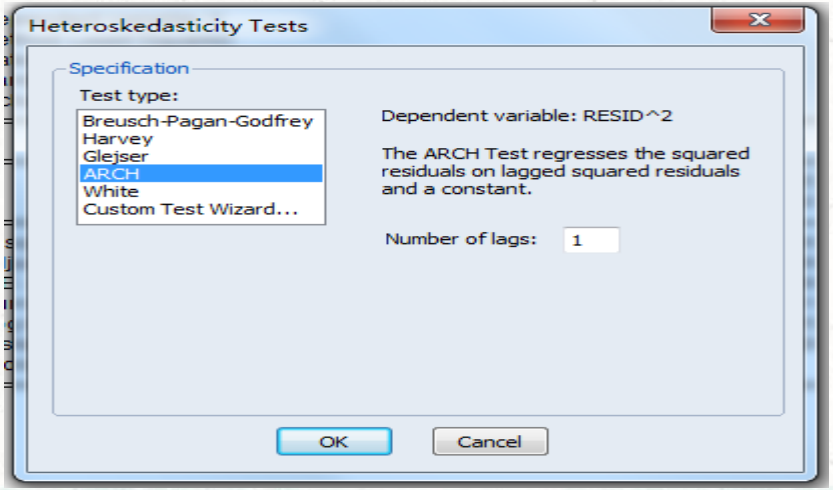


做完G-Q检验后最好重新导入数据，以免影响后续检验（ARCH）结果。

拓展：时间序列数据的情形

2. ARCH检验

在OLS估计结果界面，按路径“View/ Residual Diagnostics/ Heteroskedasticity Tests…”，在弹出的specification对话框中的检验类型中“Test type”选择“ARCH”检验：



在对话框中“Number of lags”输入相应的数值，单击“OK”即可。

滞后阶数的选择可以指定一个适当的最大阶数，然后根据AIC和SIC值选择一个最优的阶数（值越小越好）。

ARCH检验结果

本例中，取滞后阶数为1，即=1，则有 $\hat{e}_t^2 = \hat{\alpha}_0 + \hat{\alpha}_1 e_{t-1}^2$

注意：这里的obs*R-squared表示$(n-p)R^2$的数值

“30”表示n-p,不是n,

取$\alpha = 0.05$，查临界值得

$\chi^2_{0.05}(1) = 3.8415$

由于$(n-p)R^2 = 11.28965 >$

或看P值 $\chi^2_{0.05}(1) = 3.8415$

则拒绝原假设，表明模型显著性地存在异方差。

Heteroskedasticity Test: ARCH

F-statistic	16.89493	Prob. F(1,28)	0.0003
Obs*R-squared	11.28965	Prob. Chi-Square(1)	0.0008

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 09/17/19 Time: 11:33
Sample (adjusted): 1979 2008
Included observations: 30 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1700.836	1113.082	1.528042	0.1377
RESID^2(-1)	0.614353	0.149465	4.110344	0.0003

R-squared	0.376322	Mean dependent var	4178.005
Adjusted R-squared	0.354047	S.D. dependent var	6377.476
S.E. of regression	5125.653	Akaike info criterion	19.98624
Sum squared resid	7.36E+08	Schwarz criterion	20.07966
Log likelihood	-297.7937	Hannan-Quinn criter.	20.01613
F-statistic	16.89493	Durbin-Watson stat	1.628291
Prob(F-statistic)	0.000312		

AIC和SIC

拓展：多元回归的情形

1. G-Q检验

可以按不同的解释变量分别排序，来进行G-Q检验，结论可能有所不同，只要有一个拒绝原假设，就认为有异方差。

2. 加权最小二乘

权重既可以尝试每一个解释变量的前面三种形式，也可以尝试多个解释变量的组合，如“1/(X2+X3)”等。

3. 对数变换法

无论是横截面还是时间序列数据，一元还是多元回归，均可以尝试用对数变换修正异方差，但记得同样要再做异方差检验，另外注意经济含义。

作 业

本科教材练习题5.3和5.4

（P132-133)