第5章　异方差

什么是异方差?

- 球形扰动项（同方差）的假设表明，在给定解释变量的条件下，不可观测误差项ε_i的方差是不变的（不依赖于x_i）。
- 如果这一点不成立，或者说对不同的x具有不同的方差，即$\mathrm{var}(\varepsilon_\mathrm{i})$依赖于$x_i$，那么误差项就是异方差。在此情形下，误差项方差

$$\mathrm{var}(\varepsilon_\mathrm{i}) = \sigma_\mathrm{i}^2 = h(X_i)$$

  就是$\mathrm{X_i} = [x_{i2}, \ldots, x_{ik}]$的函数。

-

- “条件异方差”(conditional heteroskedasticity)，简称“异方差”(heteroskedasticity)，是违背球型扰动项假设的一种情形，即条件方差$Var(\varepsilon_i|\boldsymbol{X})$依赖于$i$，而不是常数$\sigma^2$。

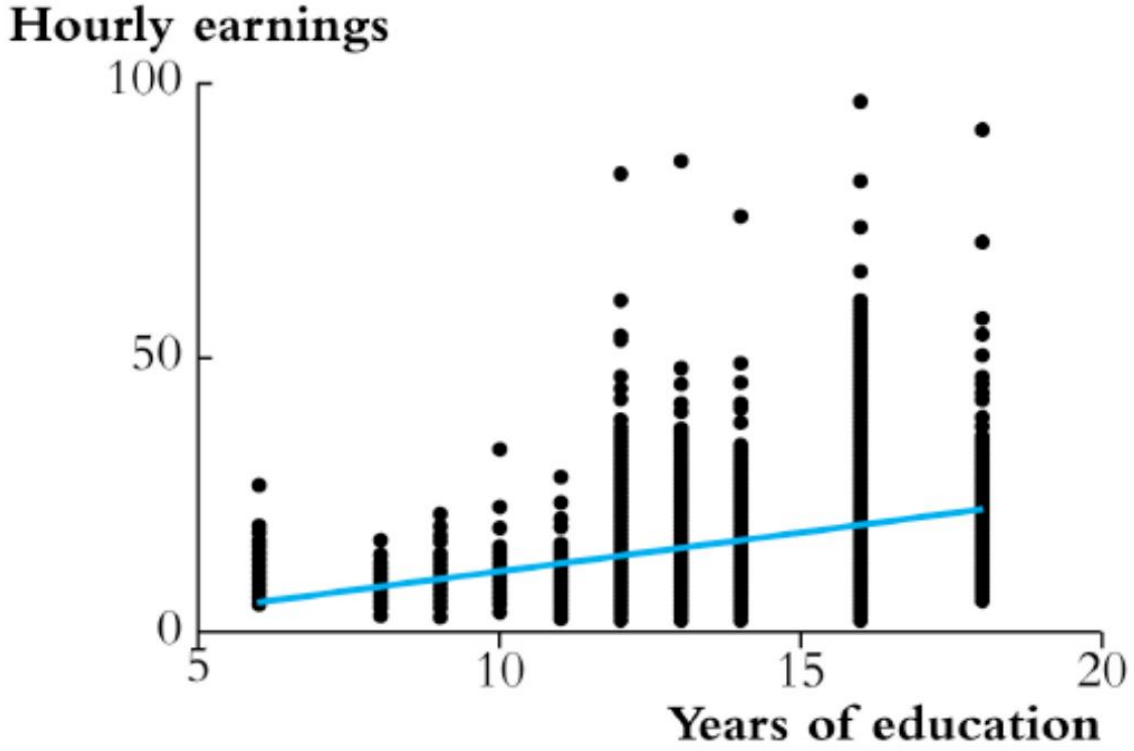
- 比如，在存在条件异方差，但无自相关的情况下，扰动项的协方差矩阵可写为

$$Var(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \begin{pmatrix} {\sigma_1}^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & {\sigma_n}^2 \end{pmatrix}$$

其中，${\sigma_1}^2, \cdots {\sigma_n}^2$不全相等

异方差的数据例子

- 从劳动经济学获得的一个真实例子——人口调查获得平均的小时工资与受教育年限的关系。



什么时候容易出现异方差?

- 通常，在使用横截面数据的时候容易出现异方差。
  - 横截面数据通常是指给定时间点众多的经济单位如企业、家庭。
  - 横截面数据总是涉及到对不同规模的经济单元的观察。
- 这意味着，随着经济单元的规模变大，与结果变量y相关的不确定性增加了。这种更大的不确定性是通过指定随着经济单元的规模越大，误差方差就越大的假定来建模的。
- 异方差并不一定是只对横截面数据的一种约束。
  - 在时间序列数据中，当我们有一个经济单元随着时间变化的观察值序列时，例如企业、家庭或整个经济体时，误差项的方差也可能会发生变化。

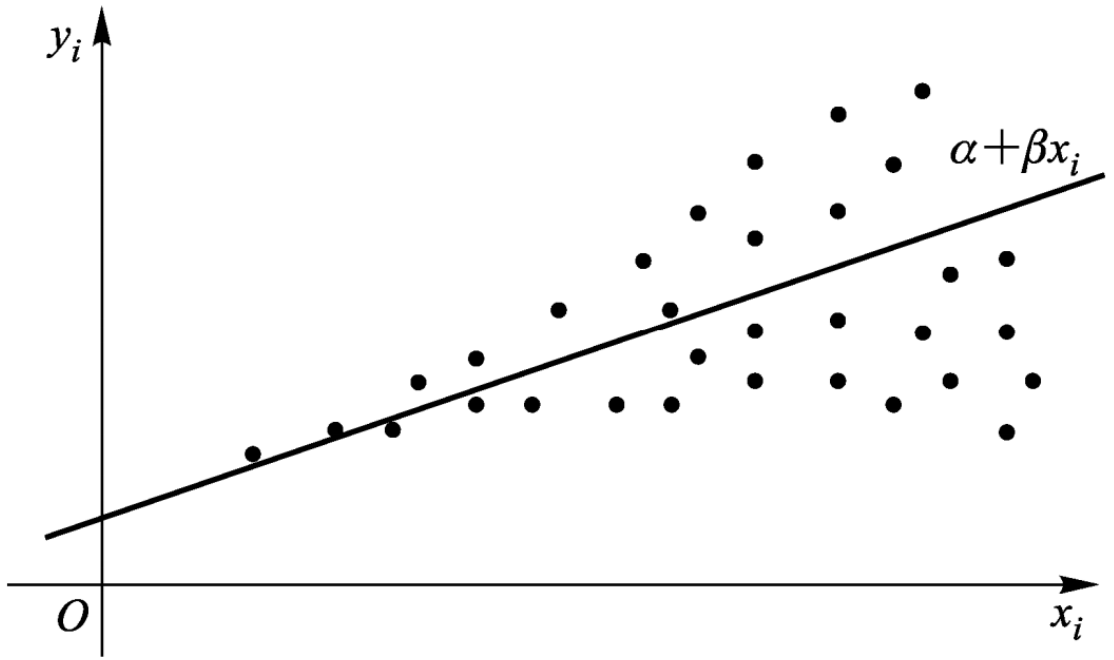
为什么要担忧异方差?

- 异方差对OLS估计量的影响:
  - OLS估计量仍然是无偏和一致的，即使我们不做同方差的假设;
  - OLS估计量尽管仍然是一个线性无偏估计量，但它不再是最优估计量。
- 如果存在异方差，那么同方差假定下推出的标准误是有偏和非一致的估计量。
  - 使用错误标准误构造的置信区间和假设检验都不可靠
  - 以前的t检验、F检验或者LM检验都不再可以用于统计推断

异方差的后果

- 在异方差的情况下：
- (1) OLS 估计量依然无偏、一致且渐近正态。因为在证明这些性质时，并未用到“同方差”的假定。
- (2) OLS 估计量方差$Var(\widehat{\boldsymbol{\beta}}|\boldsymbol{X})$的表达式不再是$\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$，因为$Var(\boldsymbol{\varepsilon}|\boldsymbol{X}) \neq \sigma^2\boldsymbol{I}$。因此，使用普通标准误的$t$检验、$F$检验失效。
- (3)高斯-马尔可夫定理不再成立，OLS不再是BLUE（最佳线性无偏估计）。

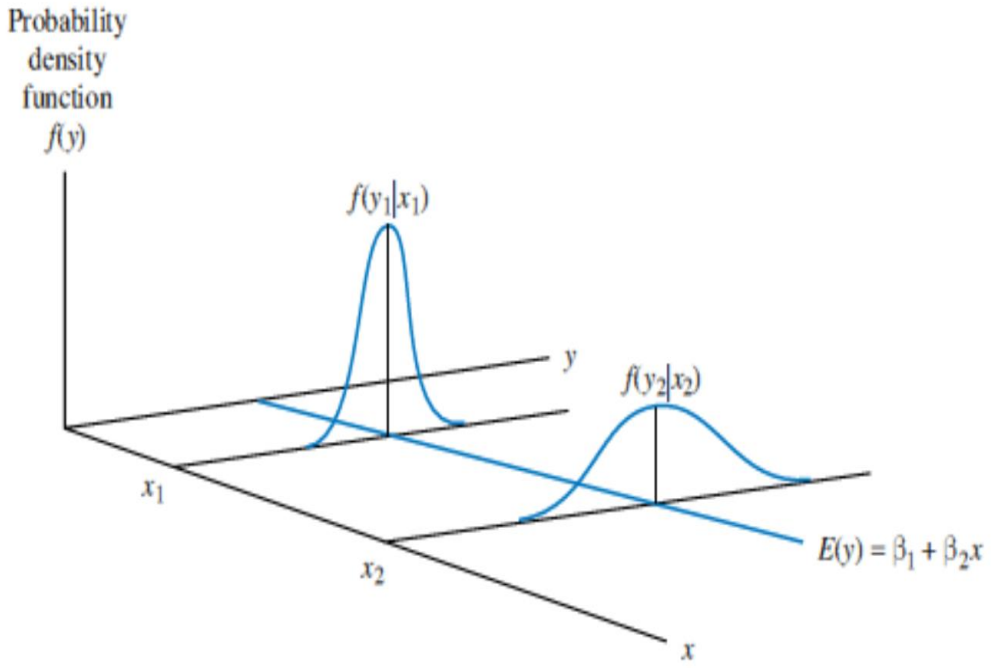
- 为了直观地理解OLS不是BLUE，考虑一元回归$y_i = \alpha + \beta x_i + \varepsilon_i$ 。
- 假设$Var(\varepsilon_i|\boldsymbol{X})$是解释变量$x_i$的增函数，即$x_i$越大则$Var(\varepsilon_i|\boldsymbol{X})$越大。



- OLS 回归线在x_i较小时可以较精确地估计，而在x_i较大时则难以准确估计。

异方差的误差项——图形

- 异方差表明y的方差依赖于x。



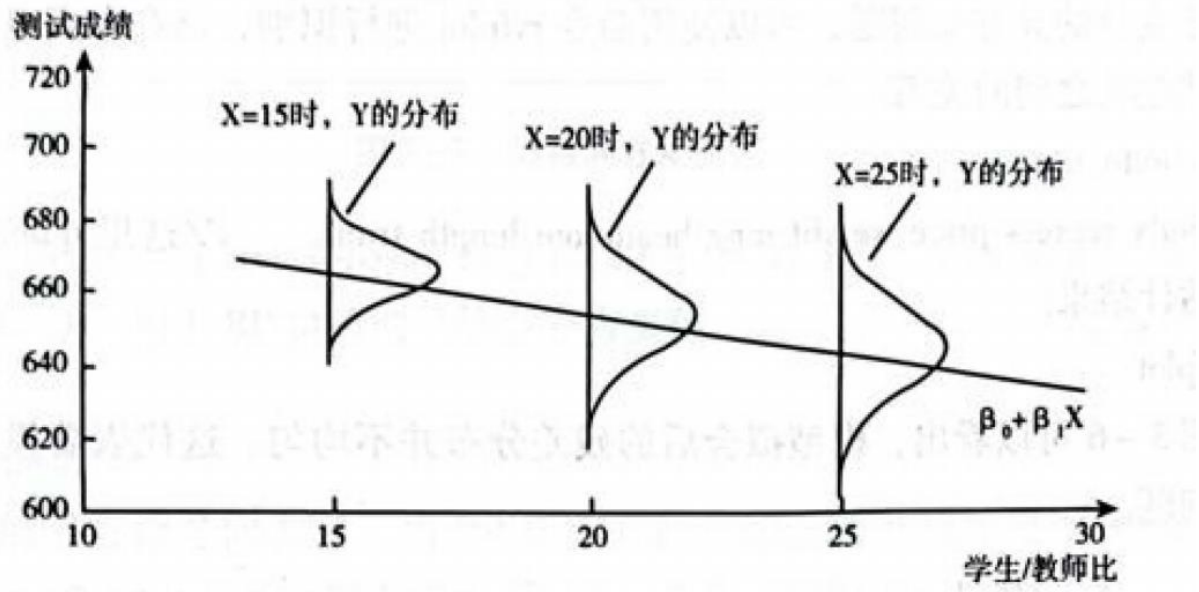


图 3－5　不同学校测试成绩条件分布

- 方差较大的数据包含的信息量较小，但 OLS 却对所有数据等量齐观进行处理；故异方差的存在使得 OLS 的效率降低。
- “加权最小二乘法”(Weighted Least Square，WLS)通过对不同数据所包含信息量的不同进行相应的处理以提高估计效率。比如，给予信息量大的数据更大的权重。

- **条件方差**(conditional variance) VS **无条件方差**(unconditional variance)
- 问题：大样本理论要求样本数据为平稳过程，而平稳过程的方差不变。大样本理论是否已经假设同方差?
- 计量经济学所指的“异方差”都是“条件异方差”，而非“无条件异方差”。
- 以一元回归模型$y_i = \alpha + \beta x_i + \varepsilon_i$为例，假设$\{x_i, y_i\}$为平稳过程，则$\varepsilon_i = y_i - \alpha - \beta x_i$也是平稳过程，故其无条件方差$\mathrm{Var}(\varepsilon_\mathrm{i}) = \sigma^2$为常数，不随$i$而变。
- 所有个体的条件方差函数$\mathrm{Var}(\varepsilon_\mathrm{i}|x_1, \dots, x_\mathrm{n})$在函数形式上也完全相同；比如，$\mathrm{Var}(\varepsilon_\mathrm{i}|x_1, \dots, x_\mathrm{n}) = {x_i}^2$。
- 但此条件方差函数的具体取值却依赖于x_i，故仍可存在条件异方差。比如，$\mathrm{Var}(\varepsilon_1|x_1, \dots, x_\mathrm{n}) = {x_1}^2$，$\mathrm{Var}(\varepsilon_2|x_1, \dots, x_\mathrm{n}) = {x_2}^2$，以此类推。

异方差的例子

- (1)考虑消费函数：$c_i = \alpha + \beta y_i + \varepsilon_i$
- 其中，c_i为消费，y_i为收入。富人的消费计划较有弹性，而穷人的消费多为必需品，很少变动。富人的消费支出更难测量，包含较多测量误差。$Var(\varepsilon_i|y_i)$可能随y_i的上升而变大。
- (2) 企业的投资、销售收入与利润：大型企业的商业活动可能动

辄以亿元计，而小型企业则以万元计；因此，扰动项的规模也不相同。如将大、中、小型企业放在一起回归，可能存在异方差。

- (3) 组间异方差：如果样本包含两组(类)数据，则可能存在组内同方差，但组间异方差的情形。

比如，第一组为自我雇佣者(企业主、个体户)的收入，而第二组为打工族的收入；自我雇佣者的收入波动可能比打工族更大。

- (4) 组平均数：如果数据本身就是组平均数，则大组平均数的方差通常要比小组平均数的方差小。

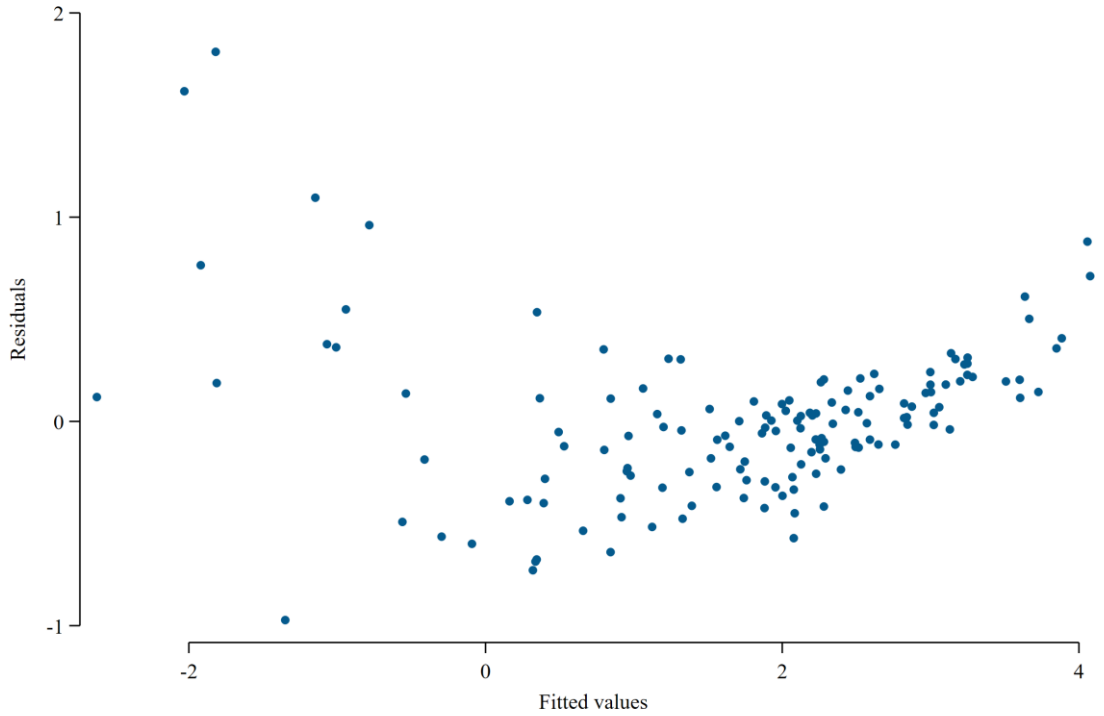
比如，考虑全国各省的人均 GDP，每个省一个数据。人口较多的省份其方差较小，方差与人口数成反比。

异方差的检验

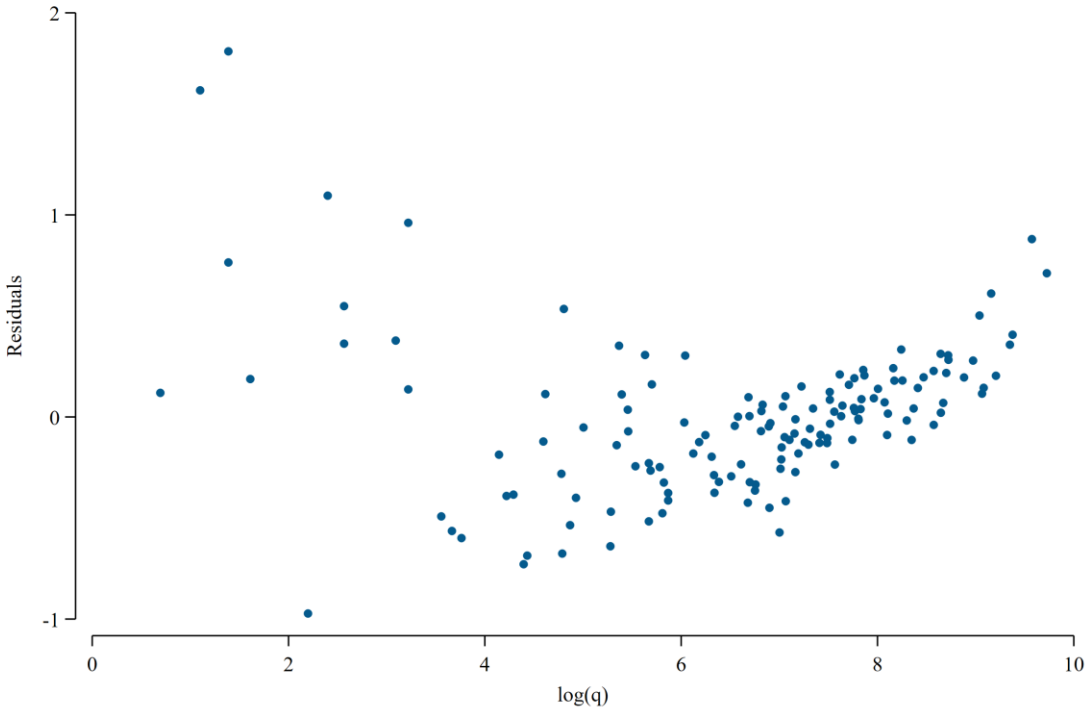
- 1.画残差图(residual plot)
- 2.BP 检验(Breusch and Pagan)
- 3.怀特检验(White)

- **1. 画残差图(residual plot)**
- 残差可视为扰动项的实现值，可通过残差的波动考察是否存在异方差。
- 可以看“残差e_i与拟合值$\hat{y}_i$的散点图”(residual-versus-fitted plot)。
- 也可看“残差 e_i与某个解释变量 x_{ik} 的散点图”(residual-versus-predictor plot)。
- 这是直观的方法，但不严格。

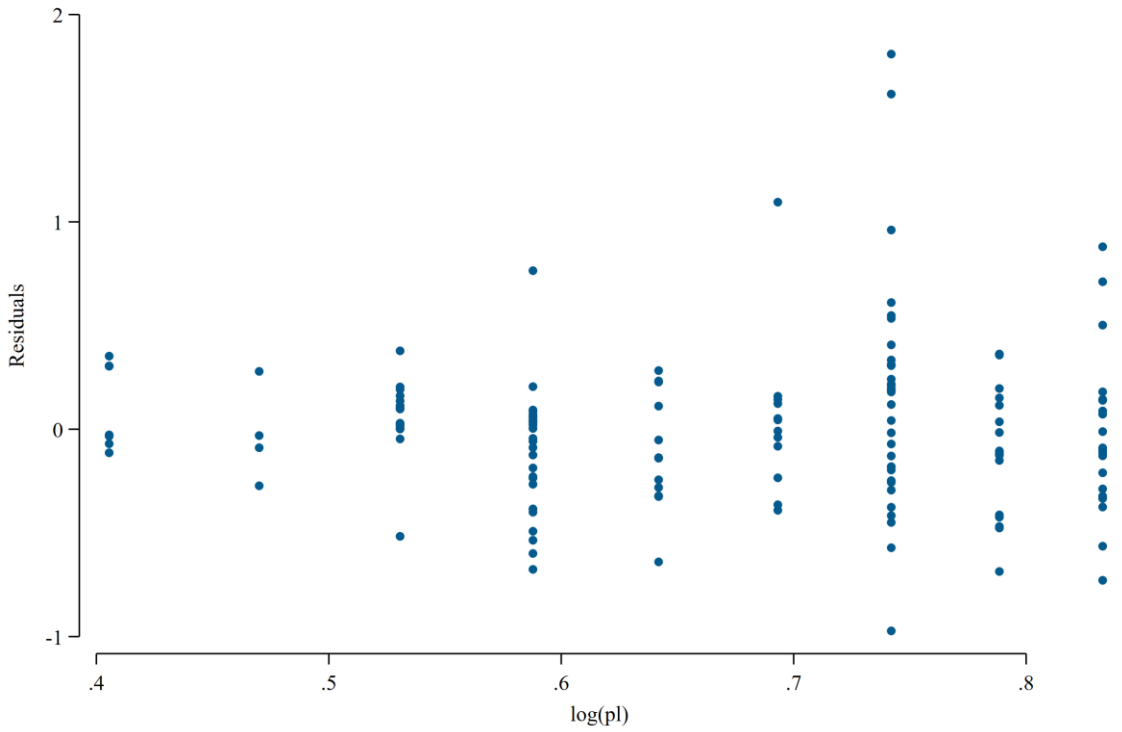
- nerlove.dta 数据集为例。
- 此数据集包括以下变量：tc (总成本)，q (总产量)，pl (工资率)，pk (资本的使用成本) 与 pf (燃料价格)，以及相应的对数值 lntc，lnq，lnpl，lnpk 与 lnpf。
- 以ols估计对数形式的成本函数



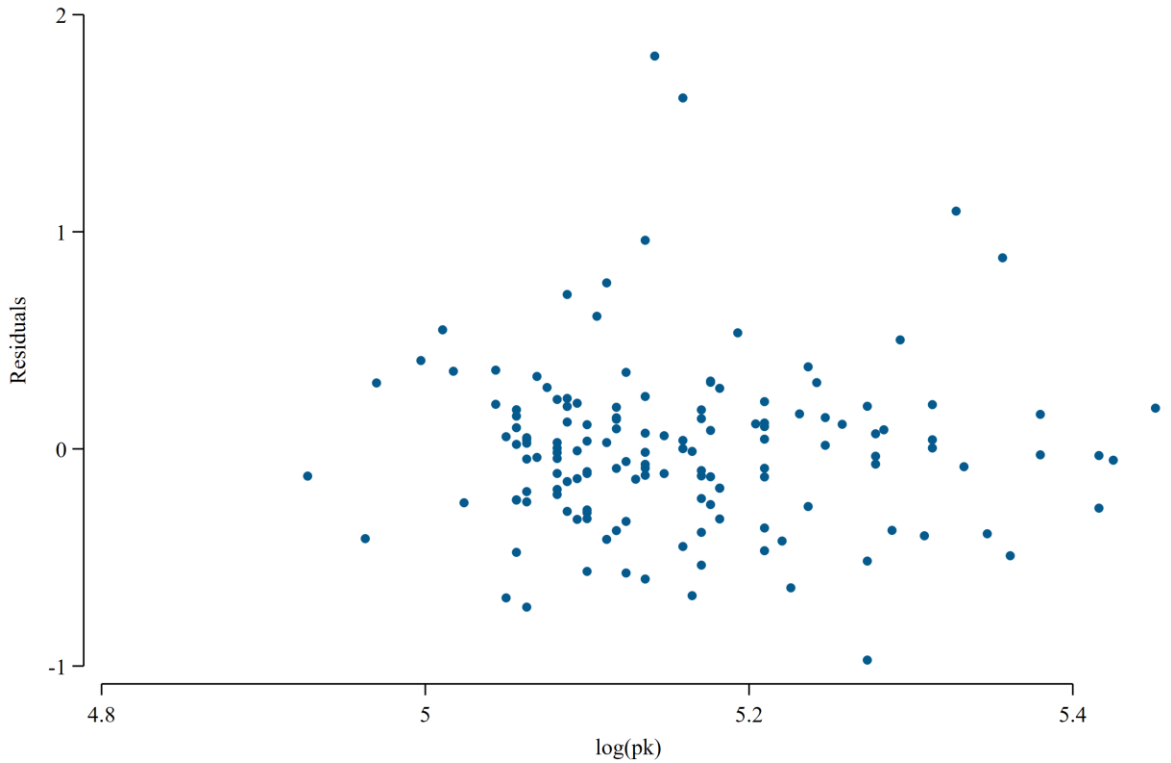
残差与拟合值的散点图



残差与解释变量lnq的散点图



残差与解释变量lnpl的散点图



残差与解释变量lnpk的散点图

- **2. BP 检验(Breusch and Pagan, 1979)**
- 假设回归模型为

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

记$\boldsymbol{x}_i = (1 \ x_{i2} \ \cdots \ x_{ik})$。

- 假设样本数据为iid，则$Var(\varepsilon_i|\boldsymbol{X}) = Var(\varepsilon_i|\boldsymbol{x}_i)$。
- “条件同方差”的原假设为

$$H_0: \ Var(\varepsilon_i|\boldsymbol{x}_i) = \sigma^2$$

- 由于$Var(\varepsilon_i|\boldsymbol{x}_i) = \mathrm{E}(\varepsilon_i^2|\boldsymbol{x}_i) - \underbrace{[\mathrm{E}(\varepsilon_i|\boldsymbol{x}_i)]^2}_{=0} = \mathrm{E}(\varepsilon_i^2|\boldsymbol{x}_i)$，原假设可写为

$$H_0: \ E(\varepsilon_i^2|\boldsymbol{x}_i) = \sigma^2$$

- 如果H_0不成立，则条件方差$E(\varepsilon_i^2|\boldsymbol{x}_i)$是$\boldsymbol{x}_i$的函数，称为“条件方差函数”(conditional variance function)。

- 假设此条件方差函数为线性函数：

$$\varepsilon_i^2 = \delta_1 + \delta_2 x_{i2} + \cdots + \delta_k x_{ik} + \mu_i$$

- 故原假设可简化为

$$H_0: \ \delta_2 = \cdots = \delta_k = 0$$

- 由于扰动项ε_i不可观测，故使用残差平方${e_i}^2$替代，进行辅助回归(auxiliary regression)：

$${e_i}^2 = \delta_1 + \delta_2 x_{i2} + \cdots \delta_k x_{ik} + error_i$$

- 记此辅助回归的拟合优度为R^2。R^2越高，则辅助回归方程越显著，越可以拒绝$H_0: \ \delta_2 = \cdots = \delta_k = 0$。

- Breusch and Pagan(1979) 使用 LM 统计量，进行 LM 检验(Lagrange Multiplier Test)：

$$LM = nR^2 \xrightarrow{d} \chi^2(K-1)$$

- 如果LM大于$\chi^2(K-1)$的临界值，则拒绝同方差的原假设。
- 为什么LM统计量是nR^2呢?
- 在大样本中，nR^2与检验整个方程显著性的F统计量渐近等价。
- 首先，对于辅助回归，检验原假设“H_0：$\delta_2 = \cdots = \delta_k = 0$”的$F$统计量为

$$F = \frac{R^2/(K-1)}{(1-R^2)/(n-K)} \sim F(K-1, n-K)$$

- 其次，在大样本情况下，F分布与χ^2分布是等价的，即

$$(K-1)F = \frac{(n-K)R^2}{(1-R^2)} \xrightarrow{d} \chi^2(K-1)$$

- 在H_0：$\delta_2 = \cdots = \delta_k = 0$成立的情况下，辅助回归方程仅对常数项回归，故
- 当$n \to \infty$时，$R^2 \xrightarrow{p} 0$，而$(1-R^2) \xrightarrow{p} 1$。
- 因此，$(K-1)F = \frac{(n-K)R^2}{1-R^2} \xrightarrow{p} (n-K)R^2$
- 在大样本下，$(n-K)R^2$与nR^2并无差别，故LM检验与F检验渐近等价。

- 如认为异方差主要依赖被解释变量拟合值$\hat{y}_i$，则可将辅助回归改为

$$e_i^2 = \delta_1 + \delta_2 \hat{y}_i + error_i$$

- 然后检验H_0：$\delta_2 = 0$(可使用 F或 LM 统计量)。
- Breusch and Pagan(1979)的最初检验假设扰动项 ε_i 服从正态分布，有一定局限性。
- Koenker (1981)将此假定减弱为 iid，在实际中较多采用。

- **3．怀特检验(White,1980)**
- BP 检验假设条件方差函数为线性函数，可能忽略了高次项。
- 怀特检验(White, 1980)在 BP 检验的辅助回归中加入所有的二次项(含平方项与交叉项)。
- 考虑以下二元回归：

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

- 除常数项外，只有两个解释变量x_{i2}与x_{i3}，故二次项包括x_{i2}^2，x_{i3}^2与$x_{i2}x_{i3}$。

- 怀特检验的辅助回归为

$$e_i{}^2 = \delta_1 + \delta_2 x_{i2} + \delta_3 x_{i3} + \delta_4 x_{i2}^2 + \delta_5 x_{i3}^2 + \delta_6 x_{i2} x_{i3} + error_i$$

- 其中，$e_i{}^2$为回归方程$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$的残差平方。
- 对原假设“H_0：$\delta_2 = \cdots = \delta_6 = 0$”进行$F$检验或$LM$检验。
- 怀特检验可检验任何形式的异方差；因为根据泰勒展开式，二次函数可很好地逼近任何光滑函数。
- 如果解释变量较多，则解释变量的二次项(含交叉项)将更多，在辅助回归中将损失较多样本容量。

异方差的处理

- 1．使用“OLS + 稳健标准误”
- 2．加权最小二乘法(WLS)
- 3．可行加权最小二乘法(FWLS)

●1．使用“OLS + 稳健标准误”

- 如发现异方差，一种处理方法是，仍进行 OLS 回归(OLS 依然无偏、一致且渐近正态)，但使用在异方差情况下也成立的稳健标准误。
- 这是最简单，也是目前通用的方法。
- 只要样本容量较大，即使在异方差的情况下，只要使用稳健标准误，则所有参数估计、假设检验均可照常进行。
- 但可能存在比 OLS 更有效率的方法，比如 WLS。

- 注释：
  - 值得注意的是，异方差稳健标准误只具有渐近大样本下的正确性
    - 因为在小样本下，通过稳健形式的标准误获得t统计量不接近于t分布，并且推断过程也不正确。
  - 术语“稳健”一词指的是在大样本下，无论误差项是同方差还是异方差，都是有效的。

实际应用中的启示

- 只考虑同方差的标准误公式和异方差稳健标准误公式一般而言是不同的——只要使用不同的公式，计算的结果一般不相同。

- 统计软件中的默认设定是同方差假定，为了获得异方差稳健标准误，必须替换这个默认设置。

- 如果不替换默认设置进行回归，那么一旦事实上是异方差，但你使用同方差假定下的标准误计算公式，将会是错误的。通常，只考虑同方差获得的标准误要小很多。

基准法则

- 无论误差项是同方差还是异方差，你都使用异方差稳健标准误形式，你都是OK的。
- 如果误差项是异方差，而你使用同方差公式计算标准误，那么标准误的计算结果将是错误的（如果存在异方差，$\hat{\beta}_i$方差的同方差估计量形式是非一致的。）
- 只有在同方差的情形下，两种公式形式才是恰好相同的。
- 基于以上的考虑，建议你总是使用上述的异方差稳健标准误形式。但是如果样本比较小，如果经过检验服从同方差，还是使用同方差吧。

在Stata中的稳健标准误

- 考虑下面的回归模型：

$$\mathrm{y_i} = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, i = 1, \ldots, N$$

- 为了在Stata中做y对$\mathrm{x_2}$和$\mathrm{x_3}$的回归和计算同方差标准误，可以输入：
- $\mathrm{reg}\ \ \mathrm{y}\ \ \ \mathrm{x_2}\ \ \ x_3$
- 为了在Stata中做y对$\mathrm{x_2}$和$\mathrm{x_3}$的回归和计算异方差一致标准误，可以输入：
- $\mathrm{reg}\ \ \mathrm{y}\ \ \ \mathrm{x_2}\ \ \ x_3, robust$
- 此时，Stata就会给出相同的OLS回归系数估计值，但是同时会报告一个新的列，称为“Robust Std. Err.”

●2．加权最小二乘法(WLS)

- 方差较小的观测值包含的信息量较大。对于异方差的另一处理方法是，给予方差较小的观测值较大的权重，然后进行加权最小二乘法估计。
- WLS 的基本思想是，通过变量转换，使得变换后的模型满足球形扰动项的假定(变为同方差)，然后进行 OLS 估计，即为最有效率的 BLUE。

- 考虑线性回归模型：

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

- 假定$Var(\varepsilon_i|\boldsymbol{x}_i) \equiv \sigma_i^2 = \sigma^2 \nu_i$，且$\{\nu_i\}_{i=1}^n$已知。
- 两边同乘权重$1/\sqrt{\nu_i}$，可得

$$\frac{y_i}{\sqrt{\nu_i}} = \beta_1 \frac{1}{\sqrt{\nu_i}} + \beta_2 \frac{x_{i2}}{\sqrt{\nu_i}} + \cdots + \beta_k \frac{x_{ik}}{\sqrt{\nu_i}} + \frac{\varepsilon_i}{\sqrt{\nu_i}}$$

- 新扰动项$\frac{\varepsilon_i}{\sqrt{\nu_i}}$不再有异方差，因为

$$Var(\varepsilon_i/\sqrt{\nu_i}) = \frac{1}{\nu_i} Var(\varepsilon_i) = \frac{\sigma^2 \nu_i}{\nu_i} = \sigma^2$$

- 对新方程进行OLS 回归，即为 WLS。
- 加权之后的回归方程满足球形扰动项的假定，故是 BLUE。

- 最小化新方程的残差平方和，可将 WLS 定义为最小化“加权残差平方和”：即

$$min \sum_{i=1}^{n} (e_i/\sqrt{v_i})^2 = \sum_{i=1}^{n} \frac{{e_i}^2}{v_i}$$

权重为$1/\sqrt{v_i}$(即方差的倒数)。

WLS 的R^2通常没有太大意义，它衡量的是变换之后的解释变量$(x_{ik}/\sqrt{v_i})$对变换之后的被解释变量$(y_i/\sqrt{v_i})$的解释力。

●3．可行加权最小二乘法(FWLS)

- 使用 WLS 虽可得到 BLUE 估计，但须知道每位个体的方差，即 $\{\sigma_i^2\}_{i=1}^n$ 。
- 实践中通常不知$\{\sigma_i^2\}_{i=1}^n$，故 WLS 事实上“不可行”(infeasible)。
- 解决方法是先用样本数据估计$\{\sigma_i^2\}_{i=1}^n$，然后再使用WLS，称为“可行加权最小二乘法”(Feasible WLS，简记 FWLS)。

- 在作 BP 检验时，进行如下辅助回归：

$$e_i{}^2 = \delta_1 + \delta_2 x_{i2} + \cdots + \delta_k x_{ik} + error_i$$

其中，$e_i{}^2$为原回归方程的残差平方。

- 通过辅助回归的拟合值，可以得到σ_i^2的估计值：

$$\hat{\sigma}_i{}^2 = \hat{\delta}_1 + \hat{\delta}_2 x_{i2} + \cdots + \hat{\delta}_{\mathrm{k}} x_{i\mathrm{k}}$$

- 但可能出现$\hat{\sigma}_i{}^2 < 0$的情形，而方差不能为负。

- 为保证$\hat{\sigma}_i^2$始终为正，假设条件方差函数为对数形式：

$$\ln e_i^2 = \delta_1 + \delta_2 x_{i2} + \cdots \delta_k x_{ik} + error_i$$

- 对此方程进行OLS回归，可得$\ln e_i^2$的预测值，记为$\ln\hat{\sigma}_i^2$。
- 得到拟合值$\hat{\sigma}_i^2 = \exp(\ln\hat{\sigma}_i^2)$（一定为正）。
- 以$1/\hat{\sigma}_i^2$为权重对原方程进行WLS估计。
- 记此估计量为$\widehat{\boldsymbol{\beta}}_{FWLS}$。

- 4．究竟使用“OLS + 稳健标准误”还是 FWLS
- 理论上，WLS 是 BLUE。
- 实践中的 FWLS 并非线性估计，因为权重$1/\hat{\sigma}_i^2$也是$\boldsymbol{y}$的函数。
- 由于$\widehat{\boldsymbol{\beta}}_{FWLS}$是$\boldsymbol{y}$的非线性函数，一般有偏
- $\widehat{\boldsymbol{\beta}}_{FWLS}$无资格参加BLUE的评选。
- FWLS 的优点主要体现在大样本中。如果$\hat{\sigma}_i^2$是σ_i^2的一致估计，则 FWLS 一致，且在大样本下比 OLS 更有效率。
- FWLS 的缺点是必须估计条件方差函数$\hat{\sigma}_i^2(\boldsymbol{x}_i)$，而通常不知道条件方差函数的具体形式。
- 如果该函数的形式设定不正确，根据 FWLS 计算的标准误可能失效，导致不正确的统计推断。

- 使用“OLS + 稳健标准误”的好处是，对回归系数及标准误的估计都一致，不需要知道条件方差函数的形式。
- 在 Stata 中操作也十分简单，在命令 reg 之后加选择项“robust”即可。
- “OLS + 稳健标准误”更为稳健(适用于一般的情形)，而FWLS 更有效率。
- 必须在稳健性与有效性之间做选择。
- 前者相当于“万金油”(谁都适用)，而后者相当于“特效药”。
- 由于“病情”通常难以诊断(无法判断条件异方差的具体形式)，故特效药可能失效，甚至起反作用。
- 如果对σ_i^2估计不准确，则 FWLS 即使在大样本下也不是 BLUE，其估计效率可能还不如 OLS。

- Stock and Watson (2012)推荐，在大多数情况下应使用“OLS + 稳健标准误”。
- 但Wooldridge(2009)指出，如果存在严重的异方差，可通过FWLS提高估计效率。
- 如果对于条件异方差函数的具体形式没有把握，不知道经过加权处理之后的新扰动项$\varepsilon_i/\sqrt{v_i}$是否同方差，可在 WLS 回归时仍使用异方差稳健标准误，以保证 FWLS 标准误的有效性。
- 如果被解释变量取值为正，有时将被解释变量取对数，可以缓解异方差问题。

谢谢！