2022 春 计量经济学复习纲要

Karry Ran

题型：

一、名词解释 2*5

- 最小二乘法
- R2 （拟合优度）
- 异方差
- 多重共线性
- 序列相关

【选择题中必然会考察到：异方差、多重共线、序列相关的】

定义 + 后果 + 产生原因 + 检测方法+ 消除方法

二、单选 10 * 2

- 一元、二元 OLS 的性质
- T 检验思路、F 检验思路

三、多选 6*3

四、证明和计算 10

- 证明一元线性回归的系数值（两种方法）、无偏性、有效性会说明
- 会估计 u 以及其方差 + beta1 的方差证明
- SST = SSE + SSR
- 异方差 序列相关 给 h(x) 去纠正一下
- LM 卡方检验 BP 检验（必考点）
- DW 统计量的计算（必考点）

五、不需要做

六、案例分析 42（主要参考期中复习的题）

- 各种计算: 调整 R2、会估计方差、t 统计量、F 统计量（思路）
- 遗漏变量的影响

1 二元回归

1. 什么是最小二乘法？【名词解释】

对于一组样本观测值（X_i, Y_i） 要找到一条样本回归线，使其尽可能地拟合这组观测值，换句话说就是使被解释变量的估计值与观测值在总体上最为接近。

2. 二元线性回归最小二乘法解的推导【必考证明题】

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$\underset{(\hat{\beta}_0,\hat{\beta}_1)}{Min}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2=\sum_{i=1}^{n}(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)^2$$

$$\Rightarrow\begin{cases}\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_i\right)=0\\ \sum_{i=1}^{n}x_i\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_i\right)=0\end{cases}\Leftrightarrow\begin{cases}n^{-1}\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_i\right)=0\\ n^{-1}\sum_{i=1}^{n}x_i\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_i\right)=0\end{cases}$$

$\rightarrow(\hat{\beta}_0,\hat{\beta}_1)$：$\hat{\beta}_1=\dfrac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}=\dfrac{\sum x_iy_i}{\sum x_i^{\ 2}}$，$\hat{\beta}_0=\bar{Y}-\hat{\beta}_1\bar{X}$

因为 $E(u\mid X)=0$ 意味着 $\mathrm{cov}(u,X)=0$，将 $Y_i=\beta_0+\beta_1X_i+u_i$ 代入即得：

$\mathrm{cov}(Y-\beta_0-\beta_1X,X)=\mathrm{cov}(Y,X)-\beta_1VarX=0$

所以：$\beta_1=\dfrac{Cov(Y,X)}{VarX}$

然后把对应的样本矩条件代入即可得到：$\hat{\beta}_1=\dfrac{n^{-1}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n^{-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}=\dfrac{\sum x_iy_i}{\sum x_i^{\ 2}}$。

3. OLS 结果的性质【选择题】

OLS 估计值的性质：

（1）$\sum\hat{u}_i=n^{-1}\sum\hat{u}_i=0$ ：残差的和或者均值为 0；

（2）$\sum\hat{u}_iX_i=0$ ，$\sum\hat{u}_i\hat{Y}_i=0$

（3） $\bar{Y}=\hat{\beta}_0+\hat{\beta}_1\bar{X}$（样本回归线必然通过点 $(\bar{X},\bar{Y})$

（4）$\bar{Y}_i=\bar{\hat{Y}}_i$（$\hat{Y}_i$ 的平均值与 Y_i 的算术平均值相等 ）

4. 拟合优度（判定系数、可决系数）

- 定义：表示回归平方和与总离差平方和之比；反映了样本回归线对样本观测值拟合的优劣程度【名词解析】
- 公式： $R^2=\dfrac{SSE}{SST}=\dfrac{SST-SSR}{SST}=1-\dfrac{SSR}{SST}$
- 性质：在 [0, 1] 回归模型中所包含的解释变量越多，R^2 越大
- 调整 R^2【必考点】：

$$\overline{R^2}=1-\frac{SSR/(n-k-1)}{SST/(n-1)}=1-(1-R^2)\frac{n-1}{n-k-1}$$

> 就相当于 SSR 和 SST 都除了个自己的自由度
>
> 若引入的解释变量没有解释能力，那么他对残差平方和的减少就没有多大贡献，反而增加了待估计参数的个数。因此修正的R2克服了未校正的判定系数随解释变量个数增加而增大的弊端。

- 一个必须要问的问题：R^2 是越大越好吗？

> 判断估计结果是否准确的标准是高斯马尔科夫定理，即 BLUE（best linear unbiased estimator-无偏、一致、有效）性质是否成立。而 BLUE 性质是否成立与 R^2 大小没有必然联系。如果高斯马尔科夫假设不成立，估计存在偏差，此时 R^2 再大也没有意义。

5. 证明【必考证明题】

$$SST = SSE + SSR$$

$$\begin{aligned}
&\sum (y_i - \bar{y})^2 = \sum [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 \\
&= \sum [\hat{u}_i + (\hat{y}_i - \bar{y})]^2 \\
&= \sum \hat{u}_i^2 + 2\sum \hat{u}_i (\hat{y}_i - \bar{y}) + \sum (\hat{y}_i - \bar{y})^2 \\
&= \mathrm{SSR} + 2\sum \hat{u}_i (\hat{y}_i - \bar{y}) + \mathrm{SSE}
\end{aligned}$$

利用 $\sum_{i=1}^{n} \hat{u}_i = 0, \sum_{i=1}^{n} x_i \hat{u}_i = 0$

因此拟合的平均值与样本平均值：$\bar{\hat{y}} = \bar{y}$

$\Rightarrow \sum_{i=1}^{n} \hat{u}_i (\hat{y}_i - \bar{y}) = 0.$

6. 六大基本假设

- **SLR.1(线性于参数)：** 参数 β 之于变量是线性的
- **SLR.2(随机抽样)：**有一个服从总体模型方程的随机样本 $\{(x_i, y_i):i = 1, 2, n\}$ 其样本容量为 n
- **SLR.3(解释变量的样本有变异)：**x 的样本结果级{x_i, i: 1, ... n} 不是完全相同的值
- **SLR.4 零条件均值：** E(u|x) = 0
- **SLR.5 同方差假设:** Var(u | x) = σ^2 => 有效
- **SLR.6 正态分布：**假设检验

7. 高斯马尔可夫定理的含义

在假定 MLR.1~MLR.5下（高斯-马尔可夫假设），OLS 方法所估计出来的参数是 BLUE（最优线性无偏估计量）—— 无偏 + 有效

8. 无偏性证明【证明题】

> 先证明 $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) y_i}{s_x^2}, \text{在这里}$$

$$s_x^2 \equiv \sum (x_i - \bar{x})^2$$

$$\sum(x_i-\bar{x})y_i=$$
$$\sum(x_i-\bar{x})(\beta_0+\beta_1x_i+u_i)=$$
$$\sum(x_i-\bar{x})\beta_0+\sum(x_i-\bar{x})\beta_1x_i+\sum(x_i-\bar{x})u_i=$$
$$\beta_0\sum(x_i-\bar{x})+\beta_1\sum(x_i-\bar{x})x_i+\sum(x_i-\bar{x})u_i$$

$$\sum(x_i-\bar{x})=0,$$
$$\sum(x_i-\bar{x})x_i=\sum(x_i-\bar{x})^2$$

$$\beta_1s_x^2+\sum(x_i-\bar{x})u_i, \text{ and thus}$$
$$\hat{\beta}_1=\beta_1+\frac{\sum(x_i-\bar{x})u_i}{s_x^2}$$

让 $d_i=(x_i-\bar{x})$, 因此有

$\hat{\beta}_i=\beta_1+\left(1/s_x^2\right)\sum d_iu_i$, 进而

$E(\hat{\beta}_1)=\beta_1+\left(1/s_x^2\right)\sum d_iE(u_i)=\beta_1$

再证明 $\hat{\beta}_0$

由于$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$
$=\beta_0+\beta_1\bar{x}+\bar{u}-\hat{\beta}_1\bar{x}$
$=\beta_0+(\beta_1-\hat{\beta}_1)\bar{x}+\bar{u}$
故而
$E(\hat{\beta}_0)=\beta_0+E[(\beta_1-\hat{\beta}_1)\bar{x}]+E(\bar{u})$
$=\beta_0$

9. 有效性的含义【选择题】

   在所有的估计中，找不出比OLS估计得到的方差再小的了，并不是说非得很小

10. 求误差方差(必考点 一定要会求)

    首先弄清楚误差和残差的区别：误差是整体模型中的 u，残差是估计模型中的 $\hat{u}$

    定义的 $\sigma^2=E(u^2)$ 注意（这个地方的方差是直接求的平方不是传统意义上的方差？ NO! 因为 u 的均值为 0 所以只不过是没写罢了）所以要是有u那 σ^2 不就有了 但是谁会告诉你 u 呢？

    因此我们只能从样本中得到 $\hat{u}$ 然后估计 σ 得到下面的无偏估计量

$$\hat{\sigma}^2=\frac{SSR}{n-k-1}$$

11. 这样就可以得到系数的方差了（证明过程要借用推导无偏性的过程）

$$Var(\beta_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$
$$Var(\beta_0) = \frac{\sigma^2 \sum_{i=1}^{n}(X_i^2)}{\sum_{i=1}^{n}(X_i - \overline{X})^2 N}$$

$$Var(\hat{\beta}_1) = Var\left(\beta_1 + \left(1/s_x^2\right)\sum d_i u_i\right) =$$
$$\left(1/s_x^2\right)^2 Var\left(\sum d_i u_i\right) = \left(1/s_x^2\right)^2 \sum d_i^2 Var(u_i)$$
$$= \left(1/s_x^2\right)^2 \sum d_i^2 \sigma^2 = \sigma^2\left(1/s_x^2\right)^2 \sum d_i^2 =$$
$$\sigma^2\left(1/s_x^2\right)^2 s_x^2 = \sigma^2/s_x^2 = Var(\hat{\beta}_1)$$

$$Var(\hat{\beta}_0) = Var(\beta_0 + (\beta_1 - \hat{\beta}_1)\overline{x} + \overline{u})$$
$$= Var((\beta_1 - \hat{\beta}_1)\overline{x}) + Var(\overline{u})$$
$$= Var(\hat{\beta}_1\overline{x}) + Var(\overline{u})$$
$$= \overline{x}^2 Var(\hat{\beta}_1) + Var(\overline{u})$$
$$= \overline{x}^2 \frac{\sigma^2}{s_x^2} + \frac{\sigma^2}{n} = \sigma^2\left[\frac{n\overline{x}^2 + \sum(x_i - \overline{x})^2}{n\sum(x_i - \overline{x})^2}\right]$$
$$= \frac{\sigma^2}{\sum(x_i - \overline{x})^2}\frac{\sum x_i^2}{n}$$

2 多元回归【为案例分析做铺垫 应该不会单独考查】

1. 为什么要拓展到多元？【选择题】
2. 遗漏变量带来的影响【案例分析题如果考到记得写好看点】

$$\widetilde{\beta_1} = \hat{\beta}_1 + \hat{\beta}_2\hat{\delta}1$$

3.

二次函数【案例分析】

$$life = 50 + 200Y - 0.05Y^2$$

1. 什么时候取最大最小
2. 边际效应的含义【语言描述】

> 在 x_i 为 n 时，由 n 到 n + 1 变化会引起 y 变化 xxx

交互项【案例分析】

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

> 一般来说，我们在研究一个问题的时候，只关心一个系数，假如说目前我们关注 $\beta 1$ 那我们不得不再去关注一下 β_3 怎么描述呢，无非是：在其他条件不变的情况下，X_1 每变化一单位 Y 变动 $\beta_0+\beta_3 X_2$ 个单位！！

3 推断（假设检验）

3.1 t 检验【案例分析题】

1. t 变量

$$t=\frac{\hat{\beta}_i-\beta_i}{se(\hat{\beta}_i)}$$

   自由度为 n - k - 1

2. 单侧和双侧检验【不管怎么样 都和 2 比】
3. 置信区间【不管怎么样 都直接乘以 2 】

   不用写最终结果 只需要列出算式即可

3.2 F 检验

1. F 检验的含义
   - H0 是所有被联合检验的参数均为 0 ，表示被解释变量与所验证的解释变量之间不存在线性关系。
   - H1 是表示至少有一个不为零，至少有一个解释变量对被解释变量的影响是显著的。
2. F 参数的构建（SSR 和 R^2 两种）【如果考难的案例估计会涉及】

$$F=\frac{(SSR_r-SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$$

$$F=\frac{(SSR_r-SSR_{ur})/q}{SSR_{ur}/(n-k-1)}=\frac{(R_{ur}^2-R_r^2)/q}{(1-R_{ur}^2)/(n-k-1)}$$

> 注意一种特殊情况，那就是检验所有的参数是否为 0 时 分子就会分别变成了
>
> - SSE/k
> - R2/k

3.3 置信区间的求解

$$\beta+-2\times\sigma$$

4 虚拟变量

1. 经济含义

考虑一个二元工资回归方程：

$$wage = \beta_0 + \beta_1 female + \beta_2 educ + u \tag{2}$$

- 此时β_1表示拥有相同学历的劳动者，其工资的性别差异
- 此时男性和女性的工资方程分别为：

$$wage = \beta_0 + \beta_2 educ + u$$

$$wage = (\beta_0 + \beta_1) + \beta_2 educ + u$$

- 期望工资是斜率相同，但截距不同的直线！β_1为直线的截距差。
- 上述设定下，男性为基准组(base group/benchmark group)；类似的，可以选择女性为基准组。

2. 交互项的含义（截距和斜率）

上述等式(4)有另外一种表达方式：

$$\ln wage = \beta_0 + \beta_1 female + \beta_2 married + \beta_3 female \times married + \beta_4 educ + u \tag{5}$$

等式(5)中$female \times married$是虚拟变量的交互项。注意等式(5)和等式(4)的区别

- β_0仍然是基准组未婚男性的工工资状况
- β_1是未婚女性较未婚男性工资差额百分比
- β_2是已婚男性较未婚男性工资差额百分比
- $\beta_1 + \beta_2 + \beta_3$是已婚女性较未婚男性工资差额百分比
- 交互项表明，性别工资差距取决于婚姻状况

3. chow 检验 —— 男性和女性工资是否真的有差异？

其实我们最初的想法就是：F 检验嘛

虚拟变量回归还可以用于检验函数形式。原假设是，H_0：两个子样本具有相同的回归函数。

- 考虑工资回归方程：

$$\ln wage = \beta_0 + \beta_1 educ + \beta_2 age + \beta_3 age^2 + u \tag{7}$$

  需要检验的是工资方程在男性和女性两个子样本中是否一致。
- 检验的办法是引入虚拟变量$female$，构造全样本的回归方程：

$$\begin{aligned}\ln wage = {} & \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 educ \times female + \beta_2 age \\ & + \delta_2 age \times female + \beta_3 age^2 + \delta_3 age^2 \times female + u\end{aligned}$$

- 原假设表述为$H_0 : \delta_0 = \delta_1 = \delta_2 = \delta_3 = 0$，可以使用$F$检验

那为什么要引入 Chow? 其实感觉和 F 差不多，但是细想你就会发现区别

- 当待检验的方程中解释变量的个数很多时，构造全样本的回归方程非常繁琐
- 注意到无约束模型的残差平方和是两个子样本组分别估计等式(7)的残差平方和相加：$SSR_{ur} = SSR_1 + SSR_2$
- 而约束方程的残差平方和为使用全样本估计等式(7)得到的残差平方和SSR_p
- 在一个含有k个解释变量的方程中，约束个数为$k+1$个，无约束方程的自由度为$N-2(k+1)$
- 此时检验的F统计量为：

$$F = \frac{(SSR_p - SSR_1 - SSR_2) / (k+1)}{(SSR_1 + SSR_2) / (N - 2(k+1))}$$

- 该检验称为“邹至庄检验”(Chow Test)，只在同方差假设下有效！

> 如果 n_2 < k+1 那还能 chow 吗？
>
> 就不能了，那就只能退化回到常规 F 检验了

5 异方差

1. 定义：$V(u_i|x_i)$!= 常数 也就是说对于不同的样本点，残差项的方差不再是一个常数

一个一般的多元线性回归模型：

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

我们在高斯-马尔科夫定理中有假设1-假设5。其中假设5为误差项u满足同方差假定。

- 若同方差假设不满足，即$\text{var}(u_i \mid \mathbf{x}_i) = \sigma_i$，则称为异方差(heteroskedasticity)
- 异方差意味着$\text{var}(u_i \mid \mathbf{x}_i)$非常数，在通常意义上是一个关于$f(\mathbf{x}_i)$的函数
- 由于每个样本的$\mathbf{x}_i$不同，因此每个样本的$\text{var}(u_i \mid \mathbf{x}_i)$也就相应不同

2. 异方差产生的原因
    - 模型中缺少某些解释变量，从而使得残差项产生某种系统模式，而非一个常数
    - 样本数据量测量误差也是导致异方差产生的原因之一，随着数据采集技术的改进，随机干扰项的方差可能减小
3. 异方差带来的后果

异方差虽然会对统计推断产生影响，但是其后果并不非常严重：

- 异方差不影响参数估计的无偏性和一致性
- 拟合优度指标R^2和$\bar{R}^2$的解释和计算也不受异方差影响
    - $R^2 = 1 - \sigma_u^2/\sigma_y^2$，都是无条件方差，$\bar{R}^2$类似
    - 无论$\text{var}(u_i \mid \mathbf{x}_i)$是否为常数，$SSR/N$都是$\sigma_u^2$的一致估计
- 异方差影响$\text{var}(\hat{\beta})$的估计，从而影响t统计量的计算
- 异方差影响统计显著性的检验！

    - 出现异方差时参数估计不再是有效的，参数估计值的方差将不再是常数，因而也不再是最小的。

- t 检验 和 F 检验失效：用于参数显著性检验的 t 统计量是在同方差的假定下服从 t 分布的，如果没有了同方差假定，则其分布就不再是 t 分布， t 检验也就失去了意义。这是因为在方差不同的情况下，无法计算出唯一的 t 值，要么偏大要么偏小。在这种情况下，两种检验都不可靠，一般会低估存在的异方差从而夸大参数的显著性
- 模型的预测失效。参数估计量的方差随着样本观测值的变化而变化，导致预测区间变大或变小，预测功能失效。

4. 异方差检验【选择题】

- 图示法：估计完模型后，计算出每一个样本的残差平方并作图（横坐标为 X 值，纵坐标为 u）
- Goldfeld-Quandt 检验 —— 样本从小到大排序后，分段回归，对分别回归出来的残差做比得到 F 统计量，若 F = 1 的假设无法被拒绝说明没有异方差
- Park 检验 —— 对图示法进行量化，但是在量化的时候必须用给定的一个残差和自变量之间的关系式（局限性）
- Glesjer 检验 —— 同 Parker 但是关系式多了
- BP 检验 —— 同 Parker 但是关系式子变了（注意 LM 的构建）

在多元线性回归模型中：

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

检验异方差的原假设是$H_0 : \text{var}\left(u_i \mid \mathbf{x}_i\right) = E\left(u_i^2 \mid \mathbf{x}_i\right) = \sigma^2$。

- 将u_i^2看做一个新的变量，该原假设意味着u_i^2不能由$\mathbf{x}_i$解释
- 显然可以通过下列回归进行检验：

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \ldots + \delta_k x_k + v$$

检验的等价原假设为$H_0 : \delta_1 = \delta_2 = \ldots \delta_k = 0$

- 在原假设成立的情况下，显然可以构造F统计量进行检验：

$$F = \frac{R_{\hat{u}^2}^2 / k}{\left(1 - R_{\hat{u}^2}^2\right) / (N - k - 1)} \sim F_{k, N-k-1}$$

- 也可以通过LM检验构造χ^2统计来检验：

$$LM = N \cdot R_{\hat{u}^2}^2 \sim \chi_k^2$$

- 上述检验过程称为布罗施-帕甘异方差检验(Breusch-Pagan Test of Heteroskedasticity, BP Test)

- White 检验 —— 不再限制一种特定的检验方程

- 异方差的检验还可以使用怀特检验(White Test)进行，其检验的思想与BP-Test类似
- 由于无法确切知道$\text{var}(u_i \mid \mathbf{x}_i) = E(u_i^2 \mid \mathbf{x}_i)$的函数形式，White建议使用二阶Taylor展开来逼近：

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \ldots + \delta_k x_k + \delta_{12} x_1 x_2 + \delta_{kk} x_k^2 + v \quad (2$$

在此基础上也使用F检验或LM检验来检验原假设

- 由于k较大时，回归项很多，因此记

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$$

检验的回归式(2)可以转化为：

$$\hat{u}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + v \quad (3$$

- 同样构造$F_{2,N-3}$或$LM \sim \chi_2^2$来检验$H_0 : \delta_1 = \delta_2 = 0$

5. 纠正异方差的方法

- 加权最小二乘法（有局限性：只能针对已知的函数形式，可是实际生活中哪来的这么多的已知啊）
  - 当$\text{var}(u_i \mid \mathbf{x}_i)$的函数形式已知时，可以使用加权最小二乘(WLS)获得有效的参数估计
  - 考虑$\text{var}(u_i \mid \mathbf{x}_i) = E(u_i^2 \mid \mathbf{x}_i) = \sigma^2 h(\mathbf{x}_i)$情况，此时有：

$$\text{var}\left(u_i/\sqrt{h_i} \mid \mathbf{x}_i\right) = E\left(u_i^2/h_i \mid \mathbf{x}_i\right) = \sigma^2$$

因此同方差假设得到满足。

  - 将多元线性回归模型方程两边同除以$\sqrt{h_i}$，得到：

$$y_i/\sqrt{h_i} = \beta_0/\sqrt{h_i} + \beta_1\left(x_1/\sqrt{h_i}\right) + \ldots + \beta_k\left(x_k/\sqrt{h_i}\right) + \left(u_i/\sqrt{h_i}\right) \quad (7)$$

  - 令$x_0^* = 1/\sqrt{h_i}$，$x_k^* = x_k/\sqrt{h_i}$，采用OLS估计：

$$y_i^* = \beta_0 x_0^* + \beta_1 x_1^* + \ldots + \beta_k x_k^* + u_i^*$$

可以得到参数的无偏有效估计。

- 可行广义最小二乘 —— 更 general
  - 上述WLS方法基于两个假设：$h_i \equiv h(\mathbf{x}_i)$函数形式已知且不含未知参数
  - 在实际应用中，h_i的函数形式显然是未知的。不失一般性，可以假设：

$$h(\mathbf{x}) = \exp(\delta_0 + \delta_1 x_1 + \ldots + \delta_k x_k)\, v \quad (8)$$

其中v是条件期望为1的随机误差项。

  - 使用WLS需要得到$h(\mathbf{x})$的估计值，亦即需要先估计参数δ_k
  - 由于$E(u_i^2 \mid \mathbf{x}_i) = \sigma^2 h(x_i; \delta)$，取对数，结合等式(8)得到：

$$\ln u_i^2 = (2\ln\sigma + \delta_0) + \delta_1 x_{i1} + \ldots + \delta_k x_{ik} + e \quad (9)$$

  - 因此在$\hat{u}_i$已知的情况下，可以通过回归等式(9)得到$\hat{h}(\mathbf{x}_i)$，进而使用WLS

$h_i \equiv h(\mathbf{x}_i)$函数形式未知时纠正异方差的具体操作步骤如下：

1. 将y_i对$\mathbf{x}_i$做OLS回归并得到残差$\hat{u}_i$
2. 计算$\ln \hat{u}_i^2$，并拟合等式(9)的回归
3. 得到等式(9)的拟合值$\hat{g}_i$，并计算$\hat{h}_i = \exp(g_i)$
4. 以$1/\hat{h}_i$为权重，用WLS重新估计方程，得到参数β的一致有效估计

上述操作步骤称为可行广义最小二乘(Feasible Generalized Least Square, FGLS)

6 序列相关 —— 只探讨一阶序列相关

> 在时间序列数据或者面板数据下，前一期会和后一期有关联，这么一想就好像是不随机抽样了一样。

1. 定义：随机干扰项不再相互独立 存在类似于 $u_t = \rho u_{t-1} + \epsilon_i$ 的关系式
2. 产生的原因：
    - 经济现象所固有的惯性
    - 数据处理的影响：得到的数据都不是原始的数据，通过已知数据采用内插或修匀得到的数据，可能会出现序列相关性。
3. 后果和异方差性相同
4. 如何检验一阶序列相关？
    - 图示检验法

      和异方差图示检验法类似，只不过横纵坐标由 $\hat{u} - x$ 变为了 $u_i - \hat{u}_{i-1}$
    - DW 检验（必考点） —— 针对一阶序列相关 设置了 DW 统计量，其实说白了就是对图示法进行量化

      画完图后我们做如下假设：

      $$u_t = \rho u_t + \epsilon_i$$

      怎么去检验这个假设呢，构建一个 D.W. 统计量

      $$D.W. = \frac{\sum_{i=2}^{n}(\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{i=1}^{n}\hat{u}_t^2} = 2 - 2\rho$$

      分别做完差分回归后，求 D.W. 就可以 求 ρ
        - 知道 ρ 求 D.W. 或者反过来
        - 得到 D.W. 统计量后，给出 ρ 并基于此判断序列相关性——是正还是负还是没有相关性。【必考】
            - $\rho = 1, D.W. = 0$ 存在正的一阶序列相关性
            - $\rho = -1, D.W. = 4$ 存在负的一阶序列相关性
            - $\rho = 0, D.W. = 0$ 不存在一阶序列相关性

      BUG:
        - 只能检测一阶序列相关性
        - 存在无法检测的部分
    - 拉格朗日乘数检验
5. 如何纠正

> 从数学上来讲本质和异方差纠正方法相同

广义差分（关键是自相关系数 ρ 估计）

> 对于一元模型：

$$Y_t = \beta_0 + \beta_1 X_t + u_t$$

如果存在序列相关性，也即：

$$u_t = \rho u_{t-1} + \epsilon$$

那么就不能直接对一元模型进行回归了，那就要构建一个能够回归的

$$\begin{aligned} & T_t - \rho Y_{t-1} \\ = & \beta_0 + \beta_1 X_t + u_t - (\rho\beta_0 + \rho\beta_1 X_{t-1} - \rho u_{t-1}) \\ = & \beta_0(1-\rho) + \beta_1(X_t - \rho X_{t-1}) + \epsilon \end{aligned}$$

这样就可以直接回归了嘛，只不过自变量有了变化，也需要对估计出来的参数做关于 ρ 的调整，所以就要估计好 ρ

- 经验法（经验为 1）
- D.W. 统计量估计 $D.W. = 2 - 2\rho$
- 格克兰特-奥卡特迭代法
- 杜宾两步法

7 多重共线性

1. 定义：多元线性回归中其基本假设之一是各个解释变量互相独立，即不存在线性相关关系，如果某两个或多个解释变量之间出现了较强的近似相关性，并且是线性相关性，则称为多重共线性。
   - 完全多重共线性：如果存在

     $$c_1 X_1 + c_2 X_2 + c_3 X_3 + \ldots + c_k X_k = 0$$

     其中 c_i 不全为 0 ，则称为解释变量的完全共线性，也就是解释变量之间存在严格的线性关系，表明至少有一个变量可以由其他变量线性表示
   - 近似共线性：如果存在

$$c_1X_1 + c_2X_2 + c_3X_3 + \ldots + c_kX_k + v_i = 0$$

其中 c_i 不全为 0 ，v_i 为随机误差项，则成为解释变量的近似共线性

2. 产生的原因
    - 经济变量之间存在内在联系：这是产生多重共线性的根本原因

        > 的确有很多的解释变量之间存在很强的线性关系

    - 经济变量在时间上具有相关的共同趋势

        > 解释变量共同涨或者共同跌

    - 解释变量中含有滞后变量

        > 就是说模型中的解释变量里既有 x_t 又有 x_{t-1} 解释变量本身就有序列相关性，所以这样就产生了多重共线性

3. 多重共线性的后果
    - 难以区分解释变量的单独影响

        > 回归系数的方差和标准差较大，参数估计的误差增大，无法正确区分各个解释变量对被解释变量的单独影响

    - 参数估计值不稳定，模型缺乏稳定性

        > 完全共线性系数不存在，近似共线性，参数估计值方差变大

    - 参数估计量的回归系数符号有误，经济含义不合理
    - 变量的显著性检验失去意义

        > 方差太大了，容易错误不拒绝

4. 多重共线性的检验

> 其实都很符合常识

- 不显著系数法
    - R^2 很大 t 很小
    - 理论很强（理论推导该参数本该很大）但检验值很小（不显著）
    - 新引入变量后，方差增大
- 判定系数检验法
    - 依次删除解释变量，看哪个变量对 R^2 的贡献最低
    - 将 X_j 对其他解释变量做回归，看系数的 F 检验
- 相关系数法：直接检验 X_i 和 X_j 的关系，做 t 检验

- 容许度和方差膨胀因子判别法，其中R_j^2都是X_j对其他变量进行回归的 R^2
  - $1-R_j^2$
  - $\frac{1}{1-R_j^3}$

5. 多重共线性的消除
   - 先验信息法：直接设定根本不含多元回归的模型
   - 改变变量的定义形式
     - 用解释变量的相对数替代绝对数
     - 删除模型中次要的可替代的解释变量
     - 差分法（根据一般经验，增量之间的共线性一般比总量之间弱的多）
   - 增大样本容量（最好的方法）
   - 逐步回归法（不断增加解释变量，逐步设置模型）

7.1 补充点：案例分析题（就是看期中考试题）

7.1.1 教育回报率（无非是换一个情景）

F 检验一定要会描述

1. 参数值要会计算（下表中所有的值都要自己手动算一下）

```
. reg lnPrice age sum har win sep
```

Source	SS	df	MS	Number of obs	=	27
				F(5, 21)	=	20.67
Model	8.50253084	5	1.70050617	Prob > F	=	0.0000
Residual	1.72802649	21	.082286976	R-squared	=	0.8311
				Adj R-squared	=	0.7909
Total	10.2305573	26	.393482974	Root MSE	=	.28686

lnPrice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.024604	.0072852	3.38	0.003	.0094536	.0397544
sum	.6071434	.1153446	5.26	0.000	.3672711	.8470157
har	-.0036647	.0009294	-3.94	0.001	-.0055975	-.0017319
win	.0011768	.0004974	2.37	0.028	.0001424	.0022112
sep	.0101573	.0555306	0.18	0.857	-.1053249	.1256395
_cons	-7.819812	1.701863	-4.59	0.000	-11.35903	-4.280595

- R 方： SS 下的 Model 即为 SSE 下为 SSR 下为 SST

$$R^2 = 8.5025/10.2305 = 0.83109$$

- 调整的 R 方 ： SS 右面就是自由度，可以根据定义来算

$$AdjustR^2 = 1-(1-R^2)\frac{n-1}{n-k-1} = 1-(1-0.83109)\frac{26}{21} = 0.7909$$

其实还有更简单的算法（用来检验）

$$AdjustR^2 = 1-\frac{0.082287}{0.39348} = 0.7909$$

- 整体的 F 见右上角 F(5, 21)

> 这个很好理解，就是说检验所有参数均为 0 的情况，构建的 F 统计量

- Root MSE

$$RootMSE = \sqrt{SSR/(n-k-1)} = \sqrt{0.822869}$$

- 置信区间

> 就用 2 倍法则即可

7.2 2021 老师透露的例子

Case 1 小班化教学

$$ln(score) = \beta_0 + \beta_1 small + \beta_2 boy + \beta_3 white + \beta_4 abscent + \beta_5 schural + \epsilon$$

1. 问：上述模型的条件均值表达式和个别值表达式

> $ln(score|x) = ln(score) = \beta_0 + \beta_1 small + \beta_2 boy + \beta_3 white + \beta_4 abscent + \beta_5 schural$
> 条件均值表达式中的残差项会被去掉
> 个别值表达式种每一个都要加上 i 的下标

2. 根据估计结果解释所有估计系数的含义

估计系数	含义
β_1	其他条件相同时，受小班化教育的学生比非小班化教育的学生成绩高的百分比

3. 进行 t 检验说明小班化教学是否对提高成绩有积极作用(关键是会语言描述)

原假设 H_0：$\beta_1 = 0$，备择假设 H_1：$\beta_1 > 0$

4. 怀疑可能遗漏了两个重要的变量：老师的教学经验 *tchexper* 和 学生是否有免费午餐 *freelunch* 问可能会对研究产生什么样的偏误？

缺失变量	影响	原因
tchexper	高估	小班里面可能配备教学经验较为丰富的老师也就是说二者存在正相关关系，另外教学丰富的老师的学生，成绩会比较高，所以 β_1 会被高估
freelunch	低估	接收小班化教育的孩子可能家庭条件比较好，所以就不会接触到免费午餐，二者是负相关关系，另外接触到免费午餐的孩子可能家里情况差，学习成绩差，和学习成绩之间有着负相关关系，所以 β_1 被高估

5. 如果小班与常规班的分组不是随机的，而是一些分班的条件被忽略了，会对 small 的估计产生什么样的影响，如何判断是否实现了随机分组？（这是一个很有心意的故事）
6. 还有哪些需要改进的地方？
    - 异方差
    - 遗漏变量
    - 内生性等等

7.3 两次期中考试，学习答题语言

第4部分（10分）

（1）（5分）基于如下陈述给出的事实和表2中的实证结果，在你看来下面说法的结论是否合理？请解释一下。

“回归（2）没有控制天生的教学能力。要做到这一点，我得到了过去一年教师的平均教学评估数据，并将其添加到回归(2)中。Beauty（美丽）的系数下降到0.051，并且在统计上是不显著的（SE=0.079）。因此，我的结论是在回归（2）中的Beauty（美丽）系数受到遗漏变量偏差影响，并且Beauty（美丽）和课程评估的真正因果关系实际上是零。”

上述说法不正确。如果在模型（2）中控制了“过去一年教师的平均教学评估”之后，Beauty（美丽）的系数下降且统计上不显著了，并不代表Beauty（美丽）和课程评估分没有影响。因为，一旦控制了“过去一年教师的平均教学评估”，此时Beauty系数的意义也随之发生了变化， Beauty的系数代表的是：在控制“过去一年教师的平均教学评估”不变的情况下，Beauty对教师课程评分增量是否有影响？（因为去年的分数已经被固定，变化的就只有今年相对于去年的变化）所以这里的问题实际上变成了“长的帅或漂亮的老师是否可以在课程评分上取得更大的进步”？这与原来的问题不再是一回事！因此此时Beauty系数变得不显著了，表明“长的帅或漂亮的老师没有在课程评分上取得更大的进步”，仅此而已。

> 这个题目还是很重要的

3.（10 分）请解释以下两个模型中交叉项系数的含义：

1） $\ln(wage) = \beta_0 + \beta_1 Post2002 + \beta_2 college + \beta_3 Post2002 * college + u$

其中：Post2002：2002 年以后为 1，2002 年之前毕业的为 0

College：大学及以上学历为 1，其他为 0

此处交叉项的系数代表，大学及以上学历相比非大学生学历者的工资差距，在 2002（高校扩招）前后发生的变化。

2） $\ln(wage) = \beta_0 + \beta_1 female + \beta_2 college + \beta_3 female * college + u$

其中：Female：女性为 1，男性为 0

College：大学及以上学历为 1，其他为 0

此处交叉项的系数代表，大学及以上学历相比非大学生学历者的工资差距，在男性与女性人群中的不同。

这是虚拟变量的典型例子：我们关注的系数其实是上大学还是没上大学，因此核心主体一定是上大学和没上大学之间的工资差距，可能要研究在不同人群中的差异，但是要注意哪个是核心

4. （10 分） 假定我们构建了以下的回归模型来考察“代际收入流动性”，即将子代的收入水平对父母的收入水平回归：

$$inc_{child} = \beta_0 + \beta_1 inc_{parent} + u$$

其中 inc_{child} 代表子代的收入，inc_{parent} 代表父母的收入，u 是残差项。

（1）请解释系数 β_1 的含义。

系数代表父母收入与子代收入的相关性，系数越大，表明两代人之间收入越相关，也就是代际收入流动性越差。4 分。

（2）参考该模型，请写出对应的计量模型来考察“富不过三代”？

$$inc_{child} = \beta_0 + \beta_1 inc_{parent} + \beta_2 inc_{grandparent} + u$$

系数 β_2 度量祖辈与孙辈之间的收入相关性（控制了父辈收入的前提下）。β_1 越大，β_2 系数越小（接近 0），则表明“富不过三代”。6 分。

第二题的模型必须这么设，因为这样最大限度地控制了变量（控制父辈的收入相同嘛）

7.4 两次作业

1. 问：根据回归结果是否可以看出来两个变量的共线性比较高？

   关键看加入一个变量后起先另一个变量的方差变化是否大。依靠的公式为：

   $$Var(\beta^2) = \frac{\sigma^2}{SST_x(1-R_j^2)}$$

2. 会写原假设和备择假设

3. 回答：基于以上的分析，你认为ros是解释CES薪金的重要因素？是否应该将其从模型中删除？

   不管回归结果怎么样，都要认定：会。一方面，从经济角度而言，认为股票收益率影响CEO报酬是合理的；基于样本而言，估计的ros系数看起来等于0的原因可能是抽样偏误所导致；另一方面，在模型中包含ros不会造成任何损害，这取决于它与其他的自变量之间的相关关系。

4. 如果原假设是 $\beta_i = 1$ 该怎么构建 t 统计量？

   $$\frac{\beta_i - 1}{se(\beta_i)}$$

5. 改写模型的例子（很有趣）

> $$lnPrice = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u$$
>
> 如果增加一个卧室就很难保证面积不改变，那如果想要直接找到一个模型做出来一个系数可以直接表示说，增加一个 150 的 bdrms 会对价格产生什么样的影响，怎么构建？
>
> $$\theta = 150\beta_1 + \beta_2$$
> $$\beta_2 = \theta - 150\beta_1$$
>
> 所以把 β_2 用 θ 替换即可：
>
> $$\begin{aligned} lnPrice &= \beta_0 + \beta_1 sqrft - 150\beta_1 bdrms + \theta bdrms \\ &= \beta_0 + (sqrf - 150bdrms)\beta_1 + \theta bdrms \end{aligned}$$

6. 虚拟变量就一个忠告：抓住主要矛盾！！！