计量经济学记忆知识点

⌛ 倒计时

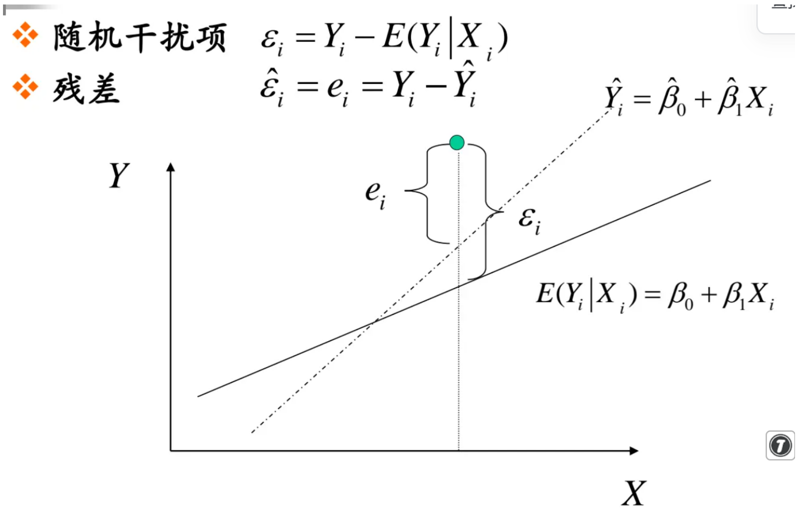
答题

1. 在保持其他变量不变的前提下
2. 假设检验记得判定符号
3. 可能存在什么问题，记得答一下当方程符号预期、显著性等（特别是多重共线性）

CH1

通过自变量在重复抽样中的已知或设定值，去估计或预测因变量的**总体均值（非具体值）**

- 研究的是 **被解释变量Y对解释变量X** 的 依赖 关系
    - 确定关系(确定变量之间的函数关系)&&统计关系(随机变量之间的**依赖**关系)【关注】
- **回归分析**的**应变量是随机的、自变量是确定的——依赖关系，不一定意味着因果关系**
    - **相关**分析两者都是随机的，不确定、不严格的依存关系
- $E(Y_i) = \beta_0 + \beta_1 X_1$ (此处没有随机扰动项！); $Y_i = E(Y_i) + \varepsilon_i$ 总体
- $\hat{Y}_i = \beta_0 + \beta_1 X_1; Y_i = \hat{Y}_i + e_i$ 样本
    - 一定通过 （$\overline{X}, \overline{Y}$） [取平均值，那么e等于0]
    - 预测值线一定通过 （$\overline{X}, \overline{Y}$） , $(X_i, \hat{Y}_i)$ [取平均值，那么e等于0]
- $\hat{\varepsilon} = e$ 总体估计值等于样本; $\overline{\hat{Y}} = \overline{Y}$



> 题目
>
> - 数据不准确可能导致回归分析的结论存在偏误
> - 回归分析考察的是解释变量与被解释变量之间的函数关系。×——是依赖关系
> - 在线性回归模型中，解释变量是原因，被解释变量是结果 错，没有因果关系

CH2

$$\hat{\beta}_1 = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{\sum x_i y_i}{\sum x_i^2}$$

残差平方和最小

β服从抽样分布，是随机变量

六个假设

假设1：回归模型是线性的，模型设定无误且含有误差项

假设2：误差项总体均值为零 $E(\varepsilon_i)=0$

假设3：所有解释变量与误差项都不相关 $Cov(X_i, \varepsilon_i)=0$

假设4：误差项观测值互不相关（无序列相关性） $Cov(\varepsilon_i, \varepsilon_j)=0$

假设5：误差项具有同方差(不存在异方差性) $Var(\varepsilon_i)=\sigma^2$

假设6：任何一个解释变量都不是其他解释变量的完全线性函数（不存在完全多重共线性）

经典线性回归模型(CLRM)的基本假设

- 估计量 $\hat{\beta}$ 服从抽样分布
- 方差估计

可用证明，随机误差项ε的方差σ^2的无偏估计

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-k} = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-k}$$

由于σ^2是未知的，可用$\hat{\sigma}^2$替代去估计参数$\hat{\beta}_j$的标准误差。

- 最佳(方程最小，有效)线性[是指**对于被解释变量**是线性的]无偏估计
- BLUE没有一致！

OLS估计量的分布为正态分布（构造出的包含估计量的统计量才服从t分布）

拟合优度

判定系数(R^2)：

$$R^2 = \frac{\text{回归平方和}}{\text{总平方和}} = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} < 1$$

总平方和 TSS：$\sum y_i^2 = \sum (Y_i - \bar{Y})^2$

回归平方和ESS：$\sum \hat{y}_i^2 = \sum (\hat{Y}_i - \bar{\hat{Y}})^2 = \sum (\hat{Y}_i - \bar{Y})^2$

残差平方和RSS：$\sum e_i^2 = \sum (Y$

如何证明？

$TSS = ESS + RSS$。

回归平方和：**解释变量变动**所引起的被解释变量的离差的大小

RSS衡量的是随机因素影响所引起的被解释变量的离差大小

- TSS:N-1 ESS:K-1 RSS:(对应残差)：N-K

- R最大为1
- $\overline{R^2} < R^2$
- 调整的判定系数可能为负数（这里的**k含常数项**，待估计参数的个数）

$$\bar{R}^2 = 1 - \frac{\sum e_i^2/(n-k)}{\sum y_i^2/(n-1)} = 1 - (1-R^2)\frac{n-1}{n-k}$$

判定系数

$$R^2 = 1 - \frac{RSS}{TSS}$$

调整的判定系数

$$\bar{R}^2 = 1 - \frac{RSS/(n-k)}{TSS/(n-1)}$$

对两个包含的解释变量个数不同的回归模型进行拟合优度比较时，选调整的判定系数

$$\overline{R^2} = 1 - \frac{RSS/N-K}{TSS/N-1} = 1 - (1-R^2) * \frac{N-1}{N-K}$$

> 题目
>
> 5 判断 (1分) 若方程总体显著性的F检验拒绝原假设，每个斜率参数的双侧t检验都拒绝“参数为零”原假设。 得分/总分
>
> A. ✕
>
> B. ✓
>
> 正确答案：A 你没选择任何选项
>
> 解析：仅某一斜率参数拒绝原假设即可
>
> - 若采用两组样本估计同一回归方程，参数估计值的差异体现了数据的随机性。

> - 随机误差项的总体均值为0以及随机误差项与解释变量不相关保证了参数估计量的无偏性。
> - 讨论回归结果时不用花费太多时间去分析常数项的估计值，这主要依据的假设是：
>
>   A.误差项与观测值互不相关 。
>
>   B误差项具有同方差。
>
>   C所有解释变量与误差项都不相关。
>
>   **D误差项总体均值为0。**
> - 若随机误差项服从t分布，OLS估计量不再具备BLUE形式（X，分布性假设不影响BLUE）
> - 拟合优度检验的是模型对**样本观测值**（*不是解释变量的估计值更不是回归线*）的拟合程度

CH3

T检验

- 显著性水平也就是**第一类错误**的概率。
- 置信区间

$$(\hat{\beta}_i - t_{\alpha/2} se(\hat{\beta}_i), \hat{\beta}_i + \boxed{t_{\alpha/2} se(\hat{\beta}_i)})$$

  - 在利用线性回归模型进行区间预测时，随机误差项的方差越大，则(预测区间越宽，精度越低)
- **自由度↑显著性↑t↓；~~分子↑分母↓F↑~~**
  - **双侧*t***检验的临界值(绝对值)大于**单侧*t***检验的临界值(绝对值)【双侧10%=单侧5%】
- 随着样本容量的增大，t值会越来越大
- 如果通过了t检验，说明是 $\beta_{实际} \neq 0$,而不是 $\hat{\beta} \neq 0$,且说明了这个变量X对Y有线性关系
- 检验显著性的统计量是 $\frac{\hat{\beta}_j}{\sqrt{var(\hat{\beta}_j)}}$ 注意上下都是估计值

F检验

> F与调整R方同方向变换

- F统计量，这里的k不含常数项

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

k（斜率个数）而不含常数项。

$$F=\frac{ESS/k}{RSS/(n-k-1)}$$

用ESS评判，所以上面是ESS

注意第一自由度是k（斜率个数）而不含常数项。

- 如果检验偏离某个特定值，注意上面的k是H0所含的，受约束的模型中可以自由取值的

❖检验某个偏回归系数等于某一特定值

$$Y_i=\beta_0+\beta_1N_i+\beta_2P_i+\beta_3I_i+\varepsilon_i$$

$$H_0:\beta_2=0.3$$

➢受约束模型 $Y_i-0.3P_i=\beta_0+\beta_1N_i+\beta_3I_i+\varepsilon_i$

➢无约束模型 $Y_i=\beta_0+\beta_1N_i+\beta_2P_i+\beta_3I_i+\varepsilon_i$

➢***F* 统计量**

$$F=\frac{(RSS_R-RSS_{UR})/1}{RSS_{UR}/(n-4)}\sim F(1,n-4)$$

- 检验整体的时候，不检验截距项

- 如果通过了F检验，那么任意一个 β （实际值）不等于0 都拒绝
- F检验 受约束模型（上面的自由度数）是受到约束的自由度，即如果是 $\beta_1+\beta_2=0$ 那么自由度应该是1.

邹检验

- 误差项是**独立且具有同方差**的**正态分**布变量。
- 断点已知，未知改变时源于k还是截距项还是都有
- 可以检验多个变化

正态性检验

- 雅克－贝拉（Jarque－Bera）检验（H0：具备正态性）

$$JB=n[\frac{S^2}{6}+\frac{(K-3)^2}{24}]\sim\chi^2(2)$$

CH4模型设定

解释变量

- 四个
    - 理论
    - 显著性
    - 拟合优度
        - AIC,SC都是越小越好；SC的惩罚比AIC更严厉
        - **被解释变量Y被变换时，这些指标不能比较**
        - 解释变量变换的时候，可以比较。
    - 偏误

函数形式

> 题目
>
> - 因变量形式不同，不能比较调整的判定系数（如Y和lnY不能比较）

常数项

- 不能删因为
    - 会使得**斜率系数的估计值具有偏差**，被放大
    - 残差均值不一定为0（违背假设）
- 即使常数项=0，也不能删，

不同函数

> 可支配收入对消费的影响：$Y = \beta_0 + \beta_1 X_1 + \epsilon$（$\surd$）；消费量ln

- 线性
    - 斜率意义：边际效应
    - 适用于X和Y之间的斜率固定
- 对数
    - 变量（包括解释变量和被解释变量）是以自然对数
    - 适用情况：
        - 全对数：弹性（X增1%，Y增百分之几）固定、斜率不固定
        - 半对数：右半对数——存在“增长率逐渐下降”的模型；左半对数——解释变量每变化1单位，被解释变量将以百分比的形式变化**（这个时候弹性为βX）**
- 多项式
    - Y由解释变量的函数形式表示

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 (X_{1i})^2 + \beta_3 X_{2i} + \varepsilon_i$$

  - 其他解释变量保持不变，如果X增加 1单位，Y将会增加β1+β2X单位（实际意义难以解释）
- 需求和收入 ×

CH5虚拟变量

- 引入**加法**形式的虚拟变量主要考察**截距**的不同。
- 引入乘法形式的虚拟变量主要考察斜率的不同。

❖加法形式：考察截距的不同

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \mu_i$$

单独采用乘法形式的情况极少

❖乘法形式：考察斜率的不同

$$C_t = \beta_0 + \beta_1 X_t + \beta_2 D_t X_t + \mu_t$$

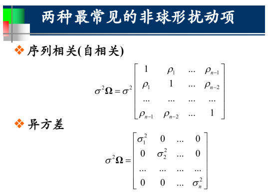
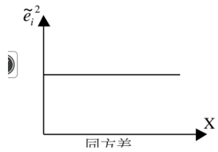
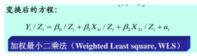
D_iX_i：被称为斜率虚拟变量

❖混合形式：截距和斜率同时发生变化

$$Y_i = \beta_0 + \beta_1 X_i + \beta_3 D_i + \beta_4 (D_i X_i) + \mu_i$$

问题解决方案一览表

存在问题	后果	来源/形式	诊断	弥补修正
遗漏变量	有偏且**非一致**，方差变小 不符合预期符号		• 原假设：不是遗漏变量	
不相干变量	**无偏**但**非有效(方差增大)**，t检验失效		• 原假设：**是**冗余变量 （备择假设希望出现的结果，唯一符合）	
多重共线性	**无偏**但**非有效(估计的方差增大)**。标准误非常大，即估计的精度很小。F 检验显著拟合优度高，单个系数的 t 检验**不显著**； （完全共线无解） > 多重共线性只影响有共线性的方差，没有的基本不影响。完全多重共线性**不带**随机误差项	时间序列样本（经济繁荣收入等都增长） 包含滞后变量 横截面数据（大企业都大，小企业都小）	• 相关系数 0.8 • VIF（把XiXi对其他解释变量进行OLS回归，>5） $VIF[\hat{\beta}_i] = (1 - R_i^2)^{-1}$ • t+F	• 剔除支配变量（如销售量与销售额） • 增加样本 • 剔除多余的变量 • 变换解释变量 ◦ 差分法（可能出现误差项的自相关）

				◦ 比率变换（可能带来异方差） • 无为而治
异方差 非球形	**无偏**但**非有效(估计的方差偏小，偏差通常是负)**高估了参数的t 值，变得显著 横截面数据的回归更容易产生异方差。 • 异方差检验主要关注残差（×，关注残差平方！） • 异方差性的检验思路与序列相关性的检验思路是相似的。（×，存在较大**差异**） 	非纯异方差：设定偏误导致（遗漏变量） 纯异方差：横截面数据（收入储蓄）、**时间序列相关（学习效应、犯错方差减小）**	思路：检验随机误差项**方差**与**比例因子**Z（或解释变量X）之间的**相关性**及其 “形式” • 图解法（一马平川是同方差）  • Park检验 ◦ **前提是知道比例因子** ◦ Park检验采用的是**双对数**回归模型。方差为Z^2 $\ln(e_i^2)=\alpha_0+\alpha_1\ln(Z_i)+u_i$ • BP检验[只是来源于**各个解释变量**] $e_i^2=\alpha_0+\alpha_1X_{1i}+\alpha_2X_{2i}\ \ldots+u_i$ 原假设：不存在异方差 $n\cdot R^2 \underset{asy}{\sim} \chi^2(df), df=k-1$(辅助回归斜率系数个数) R是调整前 • white检验：异方差来源于解释变量**及其高次方** $e_i^2=\alpha_0+\alpha_1X_{1i}+\alpha_2X_{2i}+\alpha_3X_{1i}^2+\alpha_4X_{2i}^2+\alpha_5X_{1i}X_{2i}+u_i$ **White检验中可包含解释变量的交叉项，也可不包含解释变量的交叉项。**	• 广义加权最小二乘法：加权最小二乘（*知道比例因子Z*）WSL  ◦ 需要知道比例因子 • **逐步回归法** • 修正异方差的标准误：White方法、 ◦ 只修正标准误而不会改变系数的估计值 ◦ 在大样本估计中，效果更好 • 重新定义新变量（推荐）：双对数
序列相关性 非球星	**无偏**但**非有效(估计的方差偏小，偏差通常是负)**高估了参数的t 值，变得显著	一阶序列相关：与其**滞后一期**值有关	思路：检验随机误差项*（没有平方啦）*之间的**相关性**及其 “形式”	非纯（设定）：特别是出现负的序列相关时，更应选择正确的

> 是随机误差项序列相关，而不是X，相关那么COV≠0

高阶序列相关：多期

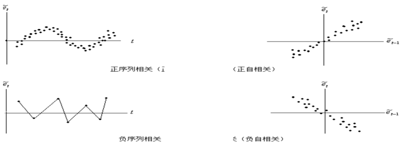
$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

➢ 其中，ρ 为一阶自相关系数，且 $|\rho|<1$；u_t 为白噪声序列，

➢ $\rho > 0 \Rightarrow$ 正序列相关
➢ $\rho < 0 \Rightarrow$ 负序列相关
➢ $\rho = 0 \Rightarrow$ 无序列相关

非纯序列相关：**设定**偏误（**遗漏变量+函数设定**）

❖ 图解法(非正式方法)
➢ 将残差对时间描点，以发现残差在时间上的特定关联
➢ 还可以怎样描点画图?



- DW（K不包含解决项，需要N样本数）
  - 要求：包含截距、一阶，不能把滞后**被解释变量**作为**解释变量、误差项非正态分布不可靠（要求:德宾-沃森d检验假定误差项具有同方差性。）**

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \quad = 2(1-\hat{\rho})$$

原假设 $\rho = 0$ 即不存在序列相关。（注意**正→负**）

正自相关	不确定区	不能拒绝无自相关	不确定区	负自相关

0 $\quad d_L \quad d_U \quad 4-d_U \quad 4-d_L \quad 4$

  - 也可进行单侧检验！备择假设：希望出现的结果

➢ 检验序列正相关 $H_0: \rho \le 0; H_A: \rho > 0$

拒绝原假设	不确定	不能拒绝原假设

$0 \quad d_L \quad d_U$

- BG检验/LM检验
  - 可高阶可滞后

✓ 针对回归模型 $Y_t = \beta_0 + \beta_1 X_{1t} + \ldots + \beta_k X_{kt} + \varepsilon_t$
✓ 假设干扰项存在p 阶序列相关

$$\varepsilon_t = \rho_1\varepsilon_{t-1} + \rho_2\varepsilon_{t-2} + \ldots + \rho_p\varepsilon_{t-p} + u_t$$

✓ 检验原假设

$$H_0: \rho_1 = \rho_2 = \ldots = \rho_p = 0$$

✓ Step2：辅助回归(将残差对解释变量、残差滞后值进行回归)获得R2

$$e_t = \alpha_1 + \alpha_2 X_{1t} + \ldots + \alpha_{k+1} X_{kt} + \rho_1 e_{t-1} + \rho_2 e_{t-2} + \ldots + \rho_p e_{t-p} + \nu$$

✓ Step3：构造LM统计量

$$LM = (n-p)R^2 \sim \chi^2(p)$$

注意是服从p阶

- 广义差分法GLS
  - 自相关系数 ρ 未知时，需要首先估计自相关系数
  - 然后不断迭代

3. 估计ρ 的迭代方法（科克伦-奥科特 迭代法）

$$(1) Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t, (2) \varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

$$Y_t - \hat{\rho}Y_{t-1} = \beta_0(1-\hat{\rho}) + \beta_1(X_t - \hat{\rho}X_{t-1}) + u_t$$

Step1：估计模型1，获得残差 $\hat{\varepsilon}^{(1)}$；再根据残差估计 $\hat{\rho}^{(1)}$
Step2：将$\hat{\rho}^{(1)}$代入广义差分模型估计 $\hat{\beta}_1^{(2)}, \hat{\beta}_2^{(2)}$，获得 $\hat{\varepsilon}^{(2)}$ 再根据残差估计$\hat{\rho}^{(2)}$
Setp3：计算 $\hat{\rho}^{(2)} - \hat{\rho}^{(1)}$，若精度满足要求，则停止迭代；若精度不满足要求，重复第二步直至精度符合要求

  - 两阶段最小二乘法（注意这是检测序列相关，不要和异方差的广义最小二乘法搞混）
  - GLS估计量是广义回归模型中的BLUE估计量。

> GLS的DW值和调整的R2不能与OLS进行比较
>
> GLS的估计值通常与OLS的估计值不同。
>
> 小样本中，因 ρ **自相关系数**估计可能有偏，GLS估计系数可能也有偏。

- NW方法尼威一韦斯特（Newey-West HAC）方法：
  - 只修正标准误而不会改变系数的估计值

				◦ 在大样本估计中，效果更好 如果不存在序列相关，则广义差分法或NW法等价于普通最小二乘法

> 消除序列相关的一阶差分变换假定自相关系数 ρ 必须等于 1 √
>
> 广义差分法（两阶段最小二乘法） 序列相关性； **广义加权最小二乘法** 异方差；
>
> 正态性 存在

虚拟应变量

LPM

- 存在问题
  - 取值不在0-1
  - 调整判定系数拟合优度不好
  - 随机干扰项**不服从正态分布**且**存在异方差**

Logit

$$D_i = \frac{1}{1+e^{-y_i}} = \frac{1}{1+e^{-[\beta_0+\beta_1 X_i+\varepsilon_i]}}$$

- 机会比率

$$\ln\left(\frac{D_i}{1-D_i}\right) = y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

机会比率（odds ratio）

- 估计方法是最大似然法
- 统计性质大样本性质，即估计的**标准误差是渐近性**

- z统计量，McFadden的R2，似然比（LR）统计量【自由度为k，不含常数项】
- 解释
  - Pi/1-Pi=exp(0.69)=1.99受教育年限每增加1年，女性参加工作的机会是原来的1.99倍
  - 受教育年限每增加1年，女性参加工作的概率增加0.17 (即0.69*0.6*0.4)
  - 受教育年限每增加1年，女性参加工作的概率增加0.17 (即0.69*0.5*0.5)

Probit

Probit模型标准范式： $Z_i = \Phi^{-1}(D_i) = \beta_0 + \beta_1 X_i + \varepsilon_i$

Logit模型和***Probit***模型不能比较参数值

增长更快，上升更笔直

- 多元Logit模型：多种选择、无次序
- 序次Logit：多种选择，有次序
  - 债券评价
  - 舆论调查结果

公式一览表

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-k} = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k}$$

$$\hat{\beta}_1 = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{\sum x_i y_i}{\sum x_i^2}$$

$$\overline{R^2} = 1 - \frac{RSS/N-K}{TSS/N-1} = 1 - (1-R^2) * \frac{N-1}{N-K}$$

$$R^2 = \frac{回归平方和}{总平方和} = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} < 1$$

- TSS(总离差):N-1
- **ESS（回归平方和）:K-1**
- RSS:(对应残差)：N-K[应变量观测值与估计值之间的总变差]
- F统计量

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

$$F = \frac{ESS/k}{RSS/(n-k-1)}$$

k（斜率个数）而不含常数项。

用ESS评判，所以上面是ESS

注意第一自由度是k（斜率个数）而不含常数项。

https://uestc.feishu.cn/sync/R0xAdaZi9sFlYVb9yPjc7RANn3e

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

$$= 2(1-\hat{\rho})$$

总体回归函数： $E(Y|X_i) = f(X_i) = \beta_1 + \beta_2 X_i$

总体回归模型： $Y = \beta_1 + \beta_2 X_i + \mu$

样本回归函数： $\hat{Y}_i = f(X_i) = \hat{\beta}_1 + \hat{\beta}_2 X_i$

样本回归模型： $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$

- 弹性（保留线性端，对数端已经融合了）

$\beta_k \frac{X_k}{Y}$

线性-线性

$\frac{\beta}{Y}$

线性-对数

βX

对数-线性

- DW 自由度k-1，park自由度n-2，BG是p