# 问题:一：自相关的概念，产生原因，如何检验，处理的方法，产生的影响。

## 一、自相关的概念

### 1.概念

对于模型



如果随机误差项的各期望值之间存在相关关系，即协方差



这时，称随机误差项之间存在自相关性或序列相关。

随机误差项的自相关性可以有多种形式，其中最常见的类型是随机误差项之间存在一阶自相关性或一阶序列相关，即随机误差项只与它的前一期值相关：，或者，则称这种关系为一阶自相关性。一阶自相关性可以表示为



其中是与的自相关系数，是满足回归模型基本假定的随机误差项。

### 2.举例

1）时间序列数据中的自相关：由于经济活动通常具有某种连续性或持久性，自相关现象在时间序列中比较常见。比如，相邻两年的GDP增长率、通货膨胀率。又比如，某个意外事件或新政策的效应需要逐步地随时间释放出来。再比如，最优资本存量需要通过若干年的投资才能逐渐达到（滞后的调整过程）。

2）截面数据中的自相关：一般来说，截面数据不容易出现自相关，但相邻的观测单位之间也可能存在“溢出效应”，这种自相关也称为“空间自相关”。比如，相邻的省份、国家之间的经济活动相互影响（通过贸易、投资、劳动力流动等）；相邻地区的农业产量受到类似天气变化的影响；同一社区内的房屋价格存在相关性。

3）对数据的人为处理：如果数据中包含移动平均数、内插值或季节调整时，则从理论上即可判断存在自相关。

4）设定误差：如果模型设定中遗漏了某个自相关的解释变量，并被纳入到扰动项中，则会引起扰动项的自相关。

## 二、产生原因

线性回归模型中随机误差项存在序列相关的原因很多，但主要是经济变量自身特点、数据处理、变量选择及模型函数形式选择引起的。

**1.经济变量惯性的作用引起随机误差项自相关**

经济变量是对客观经济现象的反映。任何一种客观经济现象都有其历史的延续性和发展的继承性，现在的状况是在过去基础上演进而来的，过去的发展水平、速度、特征都会对现在的状况产生重要影响。同一经济变量，在前期与后续时期总存在一定的相关性，不可能互不相干。随机误差项作为模型中的一个特殊经济变量，它虽然包含的具体内容很多，不具有单一的经济含义，但它与模型中独立出现的解释变量相类似，不同观测期的取值也不可能完全互不相干，总存在一定的相关性。

**2.经济行为的滞后性引起随机误差项自相关**

在时间序列数据中，经济变量的运行往往存在着一种变化趋势，表现在时间前后期的相互关联上所形成的惯性。今年的农业收成会对明年、后年或以后若干年的经济发展带来直接或间接的影响。今年的家庭消费水平在很大程度上要受到上一年消费水平的影响。如果模型忽视这些滞后因素，滞后影响将在随机误差项中体现为系统性。

**3.一些随机偶然因素的干扰引起随机误差项自相关**

在时间序列中，某一时期发生的随机冲击往往要延续若干时期。通常随机偶然因素是指战争、自然灾害、金融危机等，这些随机因素的影响不仅对当期经济活动产生影响，而且影响以后若干时期，反映在模型中很容易形成随机误差序列的自相关。

**4.模型设定误差引起随机误差项自相关**

有时自相关的发生并不是因为连续观察值之间相关，而是由于回归模型没有正确设定。模型设定不当意味着模型中遗漏了本应包括在模型中的重要变量，或添加了多余的解释变量，或模型选择了不正确的函数形式。如果发生了这样的模型设定错误，则从不正确的模型中得到的残差会呈现自相关。

**5.观测数据处理引起随机误差项自相关**

在经济计量分析中，所使用的时间序列数据，因多种原因在代表性上存在某些缺陷，为增强数据的代表性或弥补其他方面的缺陷，往往需要对原始观测数据进行内插或平滑处理。经过这样处理后的时序资料与原始观测的时间序列之间的差异便会在随机误差项中反映出来，并引起随机误差项的序列相关。

## 三、如何检验

1.画图

可以将残差*et*与滞后残差*et-1*画成散点图，也可以画残差的“自相关图”，显示各阶样本自相关系数或偏自相关系数。此法直观但不严格。

2.BG检验

原假设H0：或（一阶或多阶自相关的自相关系数=0）

检验统计量：LM统计量

3.Box-Pierce Q检验

定义残差的各阶样本自相关系数为

检验统计量：“Box-Pierce Q统计量”

经过改进的“Ljung-Box Q统计量”

两种Q统计量在大样本下是等价的，但Ljung-Box Q统计量的小样本性质更好。

4.DW检验

缺点：只适用于检验一阶自相关（当DW值接近于2时，只能说明*ut*与*ut-1*不相关，并不同时意味着模型不存在高阶自相关性，即不能得出“不存在自相关性”的结论）；必须在解释变量满足严格外生性的情况下才成立；DW统计量依赖于数据矩阵X，无法制成统计表，而必须使用上限分布*dU*与下限分布*dL*来判断，且得到*dU*与*dL*的临界值后，仍然存在无结论区域。

检验统计量：

为残差的一阶自相关系数。当DW=2时，，无一阶自相关；当DW=0时，，一阶正自相关；当DW=4时，，一阶负自相关；

## 四、处理的方法

1.使用“OLS+异方差自相关稳健的标准误(HAC)”

这种方法被称为“Newey-West估计法”，它只改变标准误的估计值，并不改变回归系数的估计值。由于HAC标准误取决于截断参数，故在实践中，建议使用不同的截断参数，以考察HAC标准误是否对截断参数的取值敏感。

2.使用“OLS+聚类稳健的标准误”

如果将观测值按聚类的归属顺序排列，则扰动项的协方差矩阵为“块对角”。此时，仍可用OLS来估计系数，但需使用“聚类稳健的标准误”。假设样本容量为N，包括M个聚类，其中第j个聚类包含Mj位个体。使用聚类稳健标准误的前提是，聚类中的个体数Mj较少，而聚类数很多；此时，聚类稳健标准误是真实标准误的一致估计。

3.使用可行广义最小二乘法（FGLS）

如果对自相关系数的估计较准确，而且满足严格外生性的假定，则FGLS比OLS更有效率。如果不满足严格外生性，而仅仅满足前定解释变量的假定，则FGLS可能是不一致的。

4.修改模型设定

如果随机误差项的自相关性是因为模型变量选择不当（遗漏了重要的解释变量或选用了无关的变量作为解释变量）所引起的，就应当对模型中包含的解释变量进行调整，去掉无关的解释变量，引入重要的解释变量。

如果是因为模型的函数形式选择失准，不能真实反映经济变量间的数量关系，引起随机误差项的自相关性，就应当舍弃原有的模型关系式，重新确定正确的模型函数形式。

## 五、产生的影响

1.OLS估计量依然是无偏且一致的，这是因为，在证明这些性质时，并未用到“无自相关”的假定。

2.OLS估计量依然服从渐近正态分布。

3.OLS估计量方差Var(b|X)的表达式不再是，因为，因此，通常的t检验、F检验也失效了。

4.高斯-马尔可夫定理不再成立，即OLS不再是BLUE。

# 问题2、异方差的概念，产生原因，如何检验，处理的方法，产生的影响。

## 一、异方差的概念

### 1.概念

古典模型假设球型扰动项。“异方差”是违背球型扰动项假设的一种情形，即扰动项方差依赖于*i*，而不是常数。

或：经典回归中的同方差是指不同随机误差项的方差相同，即。这里是一个对所有解释变量观测值都一样的常数。如果随机误差项的方差不是常数，则称随机误差项具有异方差性，即常数。

### 2.举例

一般来说，截面数据较容易产生异方差现象。

（1）消费函数：其中，C为消费，Y为收入。一般来说，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。另一方面，富人的消费支出可能更难测量，故包含较多测量误差。因此，可能随*Yi*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；如果将大、中、小型企业放在一起回归，就可能存在异方差。

（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，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。显然，人口较多的省份其方差较小，方差与人口数成反比。

（5）时间序列数据中也可能出现条件异方差（ARCH模型）。

## 二、产生原因

1.模型中遗漏了某些解释变量

如果模型中只包含所要研究的几个主要因素，其他被省略的因素对被解释变量的影响都被归入随机误差项，引起偏离均值的程度不同，可能使随机误差项产生异方差性。

2.模型函数形式的设定误差

在一般情况下，解释变量与被解释变量之间的关系是比较复杂的非线性关系。在构造模型时，为了简化模型，用线性模型代替非线性关系，或者用简单的非线性模型代替复杂的非线性关系，造成了模型关系不准确的误差。

3.样本数据的测量误差

一方面，样本数据的测量误差常随时间的推移而逐步积累，从而会引起随机误差项的方差增加。另一方面，随着时间的推移，抽样技术和其他收集资料方法的改进，也使得样本的测量误差逐步减少，引起随机误差的方差减小，从而随机项就不具有同方差性。

4.截面数据中总体各单位的差异

由于经济现象是错综复杂的，所以在实际经济现象中异方差性问题是大量存在的。被解释变量的分散幅度的变化，反映到模型中，可以理解为误差项方差的变化。例如，利用截面数据研究消费和收入之间的关系时，对收入较少的家庭，在满足基本消费支出之后剩余收入已经不多，用在购买生活必需品上的比例较大，消费的分散幅度不大。收入较多的家庭有更多可自由支配的收入，使得这些家庭的消费有更大的选择范围。由于个性、爱好、储蓄心理、消费习惯、家庭成员构成等不同造成的差异，使消费的分散幅度增大。

5.随机因素的影响

经济变量本身受很多随机因素影响（比如政策变动、自然灾害、金融危机等），不具有确定性和重复性，同时，社会经济问题涉及人的思维和行动，也涉及各阶层的物质利益，人的行为具有很多不确定因素。

## 三、如何检验

### 1.看残差图（residual plot）

具体来说，可以看“残差*ei*与拟合值的散点图”，也可以看“残差*ei*与某个解释变量*xik*的散点图”。这是最直观的方法，但不严格。

### 2.怀特检验（White test）

1）原理：条件同方差下，稳健标准误还原为普通标准误，那么二者的差别就可以用来度量条件异方差。

2）原假设*H0*：（同方差）

3）检验统计量：*nR2*

4）优点：可以检验任何形式的异方差。

5）缺点：如果*H0*被拒绝，怀特检验并不提供有关异方差具体形式的信息。

### 3.BP检验（Breusch and Pagan，1979）

1）BP检验假设条件方差函数为线性函数：

2）由于扰动项不可观测，故使用残差平方对解释变量进行辅助回归：

3）原假设

4）检验统计量：*nR2*

5）BP检验与怀特检验的区别：怀特检验还包括交互项和高次项，BP检验可以看成是怀特检验的特例。

6）BP检验的优点：具有建设性，即可以帮助确认异方差的具体形式。

## 四、处理的方法

**如果模型经过检验后发现存在异方差性，这个时候就需要采取措施对异方差问题进行修正。首先应该分析模型是否遗漏了重要的解释变量，或者模型的函数形式是否设置不当；然后考虑运用必要的估计方法，消除或减弱异方差性对模型的影响，以提高估计参数的精度。**

1.使用“OLS+稳健标准误”

该方法由怀特于1980年提出，是指先采用普通最小二乘法（OLS）估计原模型，然后用残差的平方作为相应的随机误差项方差的代表，对参数估计量的方差或标准误差进行修正。稳健标准误对于模型中可能存在的异方差问题不敏感，利用它对回归系数进行t检验和F检验都是渐近有效的。假设前提是残差项是独立分布的。只要样本容量较大，即使在异方差的情况下，若使用稳健标准误，则所有参数估计、假设检验均可照常进行。

2.广义最小二乘法(GLS)

基本思想：对违反基本假定的模型做适当的变量转换，使其转化成满足基本假定的模型，从而可以使用OLS法估计模型。

假设，其中为对称正定矩阵且已知。对于，存在非退化矩阵，使得。将原回归模型两边同时左乘矩阵C，可得。定义变量转换：。可将原模型写成，变换后的回归模型仍然满足严格外生性，且球型扰动项假定也得到满足，故高斯-马尔可夫定理仍然成立，对变换后的模型使用OLS即得到GLS估计量。

缺点：前提是必须确切知道协方差矩阵V，这在实践中并没有保证。

3.加权最小二乘法（WLS）

对原模型加权，使之变成一个新的不存在异方差性的模型，然后采用OLS法估计其参数。加权的基本思想是：在采用OLS法时，对较小的残差平方赋予较大的权数，对较大的残差平方赋予较小的权数，对残差进行校正，提高参数估计的精度。

设一元线性回归模型为，模型两端同时除以得，记，则有，此时，，原模型变成同方差模型，可以利用最小二乘法估计模型，并且得到的是最佳线性无偏估计量。

4.可行广义最小二乘法（FGLS）

先用样本数据来一致地估计*V*(*X*)，再使用GLS。

5.究竟使用“OLS+稳健标准误”还是FWLS

FWLS优点：主要体现在大样本理论中，如果是V的一致估计，则FWLS是一致的，而且在大样本下比OLS更有效。

FWLS缺点：必须估计条件方差函数，但通常并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，则根据FWLS计算的标准误可能失效，导致不正确的统计推断。

“OLS+稳健标准误”优点：它对回归系数及标准误的估计都是一致的，并不需要知道条件方差函数的形式。

总之，“OLS+稳健标准误”更为稳健（即适用于更一般的情形），而FWLS更有效。

## 五、产生的影响

1、OLS估计量依然是无偏、一致且渐近正态的。这是因为，在证明以上性质时，并未用到“同方差”的假定。

2、OLS估计量方差的表达式不再是，因为。因此，通常的*t*检验、*F*检验也失效了。

3、高斯-马尔可夫定理不再成立，即OLS不再是BLUE（BLUE：最佳线性无偏估计量）。在存在异方差的情况下，GLS才是BLUE。（*为何OLS不再是BLUE：*假设是某解释变量*xi*的增函数，即*xi*越大则越大。方差较大的数据包含的信息量较小，但OLS却对所有的数据等量齐观地进行处理。因此，从整体而言，异方差的存在使得OLS的效率降低。GLS及其特例WLS正是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，如给予信息量大的数据更大的权重。）

# 问题3、如何解决遗漏变量带来的偏差。多重共线性的概念，产生原因，如何检验，处理的方法，产生的影响。虚拟变量的概念以及它的作用，影响。

## 解决遗漏变量带来的偏差

设定模型时漏掉了一个相关的解释变量，这类错位被称为遗漏变量偏差。解决方法如下：

1、加入尽可能多的控制变量

2、使用“代理变量”：一个理想的代理变量应该满足以下条件：（1）多余性：即代理变量仅通过影响遗漏变量而作用于被解释变量。（2）剩余独立性：遗漏变量中不受代理变量影响的剩余部分与所有解释变量均不相关。

3、工具变量法

4、使用面板数据

5、随机实验与自然实验

## 二、多重共线性

### 1、概念

在一个线性回归模型中，如果某一个解释变量与其他解释变量存在线性关系（某个解释变量可以写成其他解释变量的线性组合），则称这个回归模型中存在多重共线性。

### 2、产生原因

（1）经济变量之间在时间上往往存在同方向的变化趋势

（2）经济变量之间往往存在着密切的关联度

（3）在模型中引入滞后变量也容易产生多重共线性

（4）在建模过程中由于解释变量选择不当，引起了变量之间的多重共线性

### 3、如何检验

（1）相关系数检验法

（2）法勒—格劳伯检验

（3）方差膨胀因子检验

多重共线性使参数估计值的方差增大。定义第个解释变量的“方差膨胀因子”为

表示第个解释变量与模型中其他解释变量辅助作回归模型的决定系数。度量了与其余解释变量的线性相关程度，越接近于1，就越大，说明与其余解释变量之间多重共线性越强，反之越弱。。因此，的大小反映了解释变量之间是否存在多重共线性，可用于度量多重共线性。

（4）特征值检验

（5）根据回归结果判断

### 4、处理方法

（1）如果不关心具体的回归系数，而只关心整个方程预测被解释变量的能力，或者关心具体的回归系数，但多重共线性并不影响所关注变量的显著性，则通常可以不必理会多重共线性。这是因为，多重共线性的主要后果是使得对单个变量的贡献估计不难，但所有变量的整体效应仍可以较准确的估计。

（2）如果多重共线性影响到所关心变量的显著性，则需要增大样本容量，剔除导致严重共线性的变量，或对模型设定进行修改。

（3）排除引起共线性的变量：找出引起多重共线性的解释变量，将它排除出去，以逐步回归法得到最广泛的应用。

（4）差分法：时间序列数据、线性模型：将原模型变换为差分模型。

（5）减小参数估计量的方差：岭回归法（Ridge Regression）。

（6）简单相关系数检验法

### 5、产生影响

（1）完全共线性下参数估计量不存在

（2）近似共线性下OLS估计量非有效

（3）参数估计量经济含义不合理

（4）变量的显著性检验失去意义，可能将重要的解释变量排除在模型之外

（5）模型的预测功能失效。变大的方差容易使区间预测的“区间”变大，使预测失去意义。

需要注意：即使出现较高程度的多重共线性，OLS估计量仍具有线性性等良好的统计性质。但是OLS法在统计推断上无法给出真正有用的信息。

## 三、虚拟变量

### 1、概念

虚拟变量是用来处理定性数据或分类数据，即取值为0或1的变量。构造人工变量的规则是：当某种属性存在时，人工变量的取值为1；当某种属性不存在时，人工变量的取值为0。在有常数项的模型中，如果定性指标共分M类，则最多只能有（M-1）个虚拟变量。

### 2、作用

引入虚拟变量的作用，在于将定性因素或属性因素对因变量的影响数量化。当虚拟变量取“1”值时，表示定性因素的影响发生作用；当虚拟变量取“0”值时，则表示这种属性因素的影响不发生作用。

在计量经济模型中引入虚拟变量有以下作用：

（1）可以作为定性（或属性）因素的代表，可以描述和测量定性（或属性）因素的影响。

（2）能够反映经济变量之间的相互关系，提高模型的精度。

（3）便于处理异常数据。由于某些突发因素的存在，如战争的爆发，地震、洪水等自然灾害的发生，可能会使原本比较稳定的经济关系，产生一段时间的混乱，可以利用虚拟变量对模型进行估计。

当样本资料中存在异常数据时，一般有三种处理方式：一是在样本容量较大的情况下直接剔除异常数据；二是用平均数等方式修匀异常数据；三是设置虚拟变量（将异常数据作为一个特殊的定性因素）。例如

### 3、影响

虚拟变量的引入提高了模型的解释能力，和参数的估计精度，但要注意虚拟变量的引入不当而导致的“陷阱”问题，可能造成参数无法估计。

# 问题4、工具变量的概念，能够解决什么问题。有效工具变量应该满足的条件以及有关工具变量的各种检验方式。

## 工具变量概念（IV）

某一个变量与模型中随机解释变量高度相关，但却不与随机误差项相关，那么就可以用此变量与模型中相应回归系数得到一个一致估计量，这个变量就称为工具变量，这种估计方法就叫工具变量法。

## 解决问题/适用范围

1、内生性问题产生的原因

（1）遗漏和解释变量相关的重要变量

（2）测量误差

（3）互为因果（x影响y，y也影响x）

2、解决问题

工具变量主要是应对遗漏变量，解决内生性问题。当解释变量与扰动项相关时，OLS估计量是不一致的，即无论样本容量多大，OLS估计量也不会收敛到真实的总体参数。如果产生了内生性问题，却不去解决，继续使用OLS估计，就是使得解释变量的系数估计值有偏且不一致。

3、关于工具变量的使用可以分为以下几种情况：

（1）简单回归，一个内生解释变量，一个工具变量，无其他解释变量：直接计算协方差即可；

（2）多元回归，一个内生解释变量，一个工具变量，多个外生解释变量：结构方程联立求解，其实也就是联立方程组求解；

（3）模型中只有一个内生解释变量，但是该内生变量有两个或者多个工具变量，无法取舍，并且也想充分利用两个工具变量包含的信息：2SLS, 两阶段最小二乘法；

（4）模型中有很多内生解释变量：判断阶条件，即：工具变量个数一定要大于内生变量个数（对于每一个内生变量来说），然后和3一样，使用两阶段最小二乘法即可。

## 三、有效工具变量满足条件

（1）相关性：工具变量与内生解释变量相关，即。

（2）外生性：工具变量与扰动项不相关，即。

（3）与模型中其他解释变量不相关；

（4）同一模型中需要引入多个工具变量时，这些工具变量之间不相关。

## 四、两阶段最小二乘法——提供工具变量最优线性组合（2SLS）

第一阶段：判断工具变量是否有效，因为你不能随便找一个变量就说他是工具变量，这里主要是判断是否是好的工具变量（即工具变量和内生解释变量之间的相关性，得强相关才行，如果弱相关就会出现问题），所以第一步就是用内生解释变量对工具变量和模型中的其他变量做回归。

第二阶段：用第一步中拟合的内生解释变量代入到原来的模型中，因为此时的和不相关。因为是由工具变量和其他外生变量决定的，工具变量和其他外生变量和都不相关，所以和也不相关），这个时候再进行回归，得出的内生解释变量的估计值就是无偏的。

## 五、各种检验方法

### 1、不可识别检验

使用工具变量法的前提之一是秩条件成立

原假设：

检验统计量：

秩条件成立的直观意义是工具变量与解释变量相关。因此，针对秩条件的不可识别检验也可在一定程度上验证是否存在弱工具变量，但不能取代对弱工具变量的检验。

### 2、弱工具变量检验

如果工具变量和内生解释变量相关性很小，导致工具变量法估计量的渐近方差变得很大，则工具变量中仅包含很少与内生解释变量有关的信息，利用这部分信息进行的工具变量法估计就不准确，即使样本容量很大也很难收敛到真实的参数值，这种工具变量称为弱工具变量。判断弱工具变量的方法有四种，分别为：

（1）适用“偏”：假设回归模型为，其中只有为内生解释变量。记工具变量为，其中为方程外的工具变量。在2SLS的第一阶段回归中，，其包含了内生变量与工具变量相关性的信息，但也可能由于与的相关性造成。为此，使用滤去影响的“偏”，记为。具体多低才构成弱工具变量，目前尚无共识。

（2）在第一阶段回归中，，检验原假设“”（即工具变量的系数为0）。一个经验规则是，如果此检验的F统计量大于10，则可能拒绝“存在弱工具变量”的原假设。

（3）如果假设扰动项为独立同分布，则可以使用“Cragg-Donald Wald F统计量”。

（4）如果不做独立同分布扰动项的假设，则应使用“Kleibergen-Paap Walsrk F统计量”。

解决弱工具变量问题的方法包括：1、寻找更好的工具变量 2、使用LIML，有限信息极大似然估计法。

### 3、过度识别检验

工具变量也可能和相关，导致工具变量法会产生误差，所以需要检验工具变量是否严格外生，但是这在恰好识别的时候无法做到，只有工具变量个数大于内生变量个数时才能使用过度识别检验。此检验的大前提是该模型至少是恰好识别的，即有效工具变量至少与内生解释变量一样多。在此大前提下，过度识别检验的原假设为“”。如果拒绝原假设，则认为至少某个变量不是外生的，与扰动项相关。

对所有外生解释变量与工具变量进行以下辅助回归：

原假设：

检验统计量：，为过度识别约束的个数（即多余的工具变量个数）。

### 4、该用OLS还是工具变量法：对解释变量内生性的检验

需要检验是否有内生性问题，如果没有内生性问题，那么OLS更加合适，因为OLS估计量的方差更小。

（1）Hausman test豪斯曼检验

Hausman test的原假设是：所有解释变量均为外生变量。如果成立，则OLS与工具变量法都是一致的。如果拒绝，则认为存在内生解释变量，应该使用工具变量法；反之则认为不存在内生解释变量，应该使用OLS。

缺点：它假设在成立的情况下，OLS是最有效率的。然而如果存在异方差，OLS并不是最有效的。故传统的豪斯曼检验不适用于异方差的情形。

（2）Durbin-Wu-Hausman test杜宾-吴-豪斯曼检验

该检验在异方差的情况下也适用，更为稳健。

原假设：第一阶段回归的扰动项与原模型的扰动项不相关，即对的回归系数。

如果考虑存在一个内生解释变量，对原假设进行检验；如果存在多个内生解释变量，对原假设进行检验。如果拒绝原假设，则认为存在内生解释变量；否则认为所有解释变量均为外生。

**问题5、二值选择模型和多值选择模型的概念。了解logit模型和probit模型，以及这两个模型在生物统计中的一些应用。**

**（一）二值选择模型和多值选择模型的概念**

**概念：**二值选择模型和多值选择模型都是应用于拟合离散被解释变量的回归模型，这类模型被统称为“离散选择模型”或“定性反应模型”。两种模型以被解释变量的选项个数多少进一步细分。当被解释变量为二值变量时，如考研或不考研、就业或不就业、买房或不买房等，则应用二值选择模型：用线性概率模型LPM，并以“连接函数” 来实现y的0-1二项分布： ;

当被解释变量为多值变量，如对不同交通方式的选择（走路、骑车、乘车）、对不同职业的选择等，则应用多值选择模型：对于个体i选择方案j带来的随机效用为 ，故个体i选择方案j的概率为

 .

**特点：**1、当模型的解释变量是离散变量时，线性回归方法不再适用，需要采用其他方2、二值选择模型和多值选择模型关注的是自变量的变动对因变量取值概率的影响3、实际应用中，常用Probit模型和Logit模型对离散选择模型模进行估计，两者的区别在于对连接函数具体形式的设定不同。

**（二）logit模型和probit模型**

**联系：** logit模型和probit模型是属于离散选择法模型，都是应用于离散选择模型估计过程中，为保证解释变量y满足而对模型进行不同转换得到的不同累计分布函数（cdf）。由于我们对解释变量的估计值属于0-1分布（即将理解为“”发生的概率），因为，所以连接函数的选择必须具有一定的灵活性。**区别：**1、probit模型的连接函数为标准正态的累计分布函数，即 ；而logit模型的连接函数为逻辑分布的累计分布函数，即。2、由于逻辑分布的累计分布函数有解析表达式（而标准正态分布没有），故计算logit通常比Probit更为方便，在心理学、社会学、经济学及交通领域得到了广泛的应用。3、逻辑分布的累计分布函数图形与标准正态分布的图形相似，其密度函数关于原点对称，期望为0方差为 (大于标准正态的方差)，因此逻辑分布具有厚尾的特性，更接近于自由度为7的t分布。

**（三）logit模型和probit模型在生物统计上的应用**

**一、应用模型：**

**logit模型**属于概率性非线性回归，主要研究二分类结果、多分类结果与一些一宁乡因素影响的一种多变量分析方法。由于其计算方法简单、速度较快被广泛应用于生物统计分析中，目前发展为多项logit模型、条件logit模型、混合logit模型、嵌套logit模型等模型及处理方法组成的模型体系。

**probit模型**源于生物实验的统计处理，目前已发展为一个独立的广义线性模型。由于probit模型较好的统计性质，是目前使用较为广泛的统计模型之一，且自身也发展为一个有二值probit模型、含有内生变量的probit模型、双变量的probit模型、部分可观测的双变量probit模型等模型及处理方法组成的模型体系。

**二、具体分析：**

**1、边际效应分析。**(1)平均边际效应:即分别在每个样本观测值上的边际效应然后简单算数平均；(2)样本均值处的边际效应:即在处的边际效应。(3)在某代表值处的边际效应:即给定，在处的边际效应

**2、几率比变化而分析（****）（只有Logit模型有）。在生物统计中Odd=p/(1-p)常用来说明存活与死亡的概率比，**表示解释变量增加一个微小量引起“对数几率比”的边际变化，或假设变为+1，及p为新值p\*,则新几率比与原先的几率比的比率可以写为，表示解释变量xj增加一单位引起几率比的变化倍数。计算表达式为：

**3、拟合优度（R2）：** 来衡量模型的拟合效果，其中为原模型的对数似然函数最大值，为常数项为唯一解释变量的对数似然函数的最大值。

**问题6、断尾回归的概念，断尾回归可以使用OLS， MLE估计吗？如果可以使用，需要注意什么问题。归并回归的概念，了解Tobit模型。**

**（一）断尾回归的概念：**对线性模型，假设只有满足的数据才能观测到。例：为所有企业的销售收入，而统计局只收集规模以上企业数据，比如。被解释变量在100,000处存在“左边断尾”。

**（二）断尾模型不能使用普通最小二乘估计（OLS）：**

对于正态分布，定义，则。故



对于模型，，则，

如果用OLS估计，则遗漏了非线性项，与相关，导致OLS不一致。

参见图14.3。总体回归线为，而样本回归线为。由于在处存在左边断尾，无论样本容量多大，参数估计值都无法收敛到α和β，不是一致估计。可能出现预测值yhat<=c的不可能情况。

1、忽略了一个非线性项

2、忽略了随机误差项实际上的异方差性

3、造成参数估计量的偏误，如果不了解解释变量的分布，对偏误进行估计也是十分困难。

**（三）可以使用MLE：**

断尾前的概率密度：

样本被观测到的概率：



断尾后的条件密度：



由此可以计算出整个样本的似然函数，然后使用MLE。

**（四）使用MLE应注意：**断尾回归模型最大似然估计要求其误差项服从正态分布,或至少已知误差项的边缘分布，然而估计量对误差项的分布非常敏感,当误差项的分布假定不正确时,最大似然的估计量是有偏的，另外误差项不满足方差齐性也会导致参数估计的不一致。

**（五）归并回归的概念：**对于线性模型，当 (或)时，所有都被归并为c，称为“归并数据”。例（上不封顶的数据，top coding）在问卷调查中，常有诸如“收入在￥50,000及以上”这样的选项。例（边角解）考虑买车的决定，并考察“买车开支”这个变量。如果不买车，则“买车开支”的最优解为边角解，即买车开支为0；反之，如果买车，则买车开支一定为正数。

归并回归与断尾回归不同的是，虽有全部观测数据，但某些数据的被压缩在一个点上。yi的概率分布就变成由一个离散点与一个连续分布所组成的混合分布。

**（六）Tobit模型：**Tobin (1958)提出用MLE估计归并数据，称为“Tobit”。在归并数据情况下，时的概率密度依然不变，仍为，时的分布被挤到“”上，即

该混合分布的概率密度为：

Tobit模型的**缺陷**是对分布的依赖性强，不够稳健。如果似然函数不正确（扰动项不服从正态分布或存在异方差），则QMLE估计不一致。使用Tobit模型时，**需要检验正态性与同方差性**。为了检验正态性，可将Tobit模型的MLE一阶条件视为某种矩条件，并基于此进行**“条件矩检验”**。但条件矩统计量的真实分布与渐近分布有相当差距，存在较严重的显著性水平扭曲，故使用“参数自助法”来获得校正的临界值。

为了检验同方差的原假设，考虑替代假设，其中通常等于解释变量（也可不同）。然后通过辅助回归，构建**LM统计量**来检验，如果发现扰动项不服从正态分布或存在异方差，解决方法之一为使用更稳健的“归并最小绝对离差法”。CLAD法仅要求扰动项为iid，即使在非正态与异方差情况下也一致，且在一定正则条件下，服从渐近正态分布。将归并数据模型简洁地写为：，如果；反之，，CLAD法的目标函数为离差绝对值之和：，选择β使得离差绝对值之和最小化，即可得到CLAD估计量。

**问题7、面板数据的概念，分类，主要的优点以及会带来的问题。了解面板数据中的固定效应估计量和一阶差分估量，并尝试证明这两个估计量随时点变化会发生什么变化。在处理面板数据时，应该使用固定效用还是随机效用模型。**

**（一）面板数据的概念：**在一段时间内跟踪同一组个体的数据。它既有横截面的维度（n位个体），又有时间维度（T个时期）。

**（二）分类：**1、短面板（n大T小，较多的出现在微观调查中）和长面板（n小T大，较多的出现在宏观数据中）；2、静态面板：解释变量不包含被解释变量的滞后值和动态面板：解释变量包含被解释变量的滞后值；3、平衡面板数据：每个时期在样本中的个体完全一样和非平衡面板数据：每个时期在样本中的个体不完全一样。

**（三）优点：**1、可以解决遗漏变量问题，遗漏变量常常是由于不可观测的个体差异或异质性造成的，如果这种个体差异“不随时间而改变”，则面板数据提供了解决遗漏变量问题的又一利器。2、提供更多个体动态行为的信息，例：考虑如何区分规模效应与技术进步对企业生产效率的影响。对于截面数据来说，由于没有时间维度，故无法观测到技术进步；对于单个企业的时间序列数据来说，无法区分生产效率的提高有多少是由于规模扩大，有多少是由于技术进步。但如果有面板数据就可能解决上述问题。3、样本容量较大，由于同时有截面维度和时间维度，通常面板数据的样本容量更大，从而可以提高估计的精度。

**（四）问题**：样本数据通常不满足独立同分布的假定，因为同一个体在不同期的扰动项一般存在自相关；面板数据的收集成本通常较高，不宜获得

**（五）固定效应估计量和一阶差分估计量：**

面板模型如 + 式（1）

1、对于固定效应模型，给定个体，将式1两边对时间取平均可得： +

用式（1）减去上式可得原模型的离差形式：

+

定义 ， ， ，则

由于上式中已将消去，故只要与不相关，则可以用OLS一致地估计，称为“固定效应估计量”，记为。由于主要使用了每个个体的组内离差信息，故也称为“组内估计量”。另外，为了保证与不相关，要求扰动项必须与各期的解释变量均不相关，而不仅仅是当期的解释变量。

2、对于固定效应模型，可对式1两边进行一阶差分，然后两式相减以消去个体效应（但同时也把消掉了），+

对此差分形式的方程使用OLS得到“一阶差分估计量”，记为。由于不再出现在差分方程中，只要扰动项的一阶差分 与解释变量的一阶差分不相关，则是一致的。此一致性条件比保证一致的条件更弱，这是的主要优点。

**（六）两个估计量随时点的变化**：1、固定效应估计法的优势是即使个体效应与解释变量相关，只要使用组内估计量，就可以得到一致估计。然而，在作离差转换的过程中，也被消掉了，故无法估计 ，也就是说无法估计不随时间而变的变量的影响，比如性别、种族等，这是的一大缺点；2、如果，则= 。但对于，如果为独立同分布的，则组内估计量比一阶差分估计量更有效率。因此，在实践中主要使用，而较少使用。

**（七）应如何选择固定效应或随机效应：**

H0：ui与xit，zit不相关（即随机效应模型为正确模型）。无论原假设是否成立，FE都是一致的。若H0成立，则RE比FE更有效；若H0不成立，则RE不一致。因此，若H0成立，则FE和RE估计量将共同收敛于真实参数值，即，反之，若两者差距过大，则倾向于拒绝H0。豪斯曼检验统计量：



若统计量值大于临界值，则拒绝H0

存在异方差的情况下，可通过两种方法解决：1、通过自助法，即计算机模拟在抽样的方法来计算；2、进行以下辅助回归：



然后使用聚类稳健标准误来检验原假设“H0：γ=0”这个检验在异方差的情况下也适用。

若随机效应模型成立，则OLS是一致的，因此, 接受原假设

若固定效应模型成立，则OLS不一致，即,拒绝原假设

对于非平衡面板则用代替即可。

# 了解面板数据中的固定效应估计量和一阶差分估计量，并尝试证明这两个估计量随时点变化会发生什么变化。

# 问题8：长面板数据的特点，和短面板数据的差异。对于长面板数据，样本容量较大时，可建立什么模型，简要叙述该模型。

# 长面板数据的特点，长面板和短面板数据的差异

时期长短可分为长面板和短面板（长面板往往时间跨度长，但个体数目较少，因此表现为T>N；短面板时间跨度短，个体数目较多，表现为T<N）

# 变系数模型

对于长面板数据，由于样本容量大，还可允许每位个体的回归方程斜率也不同，称为“变系数模型”。“变系数模型”分为两大类，取决于将“可变系数”视为常数还是随机变量。

1. 将可变系数视为常数

假设其中为个体i对应的系数。可对每个个体方程进行“分别回归”(separate regressions)。但如果不同个体的扰动项相关，则分别回归的效率不高。有效率的做法是，把所有个体回归方程叠放，然后使用“似不相关回归”(SUR)对整个方程系统进行系统估计。

此法的缺点是，需要估计较多参数，损失自由度。

作为折衷，可考虑“部分变系数模型”，即允许i中的部分系数(比如，研究者感兴趣的系数)依个体而变，而其余系数则不变。在此情况下，不再适用 SUR，因为各个体方程除了扰动项相关外，还拥有部分相同的系数(跨方程约束)。此时，可以使用 LSDV 法，即在回归方程中，引入个体虚拟变量，以及虚拟变量与中可变系数之解释变量的互动项。

1. 随机系数模型

将系数（斜率）视为随机变量，并假设。其中，为常数向量，而为随机向量，且满足因此，由于，通过迭代期望定律，可证明复合扰动项与解释变量不想关，故OLS一致。但复合扰动项的协方差矩阵为分块对角矩阵。Swamy (1970)提出用 FGLS 来估计此模型。

# 问题9：为什么需要分位数回归。分位数回归的估计方法是什么。

# 一、原因

1.一般的回归模型着重考察x对 y 的条件期望E(y |x )的影响，实际上是均值回归。但我们关心x对整个条件分布y | x的影响，而E(y |x )只是刻画条件分布y | x集中趋势的一个指标而已。如果y | x不是对称分布，则E(y |x )很难反映条件分布的全貌。

2.如能够估计条件分布 y | x的若干重要的条件分位数，比如中位数、1/4分位数、3 /4分

位数，能更全面认识条件分布y | x。使用 OLS 进行“均值回归”，由于最小化的目标函数为残差平方和，故易受极端值影响。

3.Koenker and Bassett(1978) 提出“分位数回归”，使用残差绝对值的加权平均作为最小化的目标函数，不易受极端值影响，较为稳健。分位数回归还能提供关于条件分布y | x的全面信息。分位数回归具有以下几个优势：首先，它对模型中的误差项不需做任何分布的假定，表现出很强的稳健性；其次，对条件分布的刻画更加细致，尤其能有效地分析数据分布中极端值的影响；还有，分位数回归通过使加权误差绝对值之和最小得到参数的估计具有大样本理论下的渐进优良性。

# 二、概念

设随机变量Y的分布函数为

对于Y的一组随机样本 样本均值是min的最优解；样本中位数是最小化残差绝对值和的解，即。



等价的表示为：；其中。

对于一般线性条件均值函数 ，通过求解得到参数估计值。而一般线性条件分位数函数为 ，通过求解得到参数估计值，对于任意的，估计称为第分位数的回归系数估计。

# 估计方法

1. 点估计

求解等价于以下线性规划问题：

其中e为单位向量。目前对上式的算法主要有如下几种：

1.单纯形算法。任选一个顶点,然后沿着可行解围成的多边形的边界搜索, 直到找到最优点。这种算法的特点决定其较适合不大的样本量和不多的变量。

2.内点法。从可行解围成的多边形一个内点出发，但不出边界，直到找到最优点。它在样本量大时效率也很高。

3.平滑算法。平滑算法在理论上比较简单，它适合处理具有大量观察值以及很多变量的数据集。

1. 区间估计
2. 直接估计法。该方法依据估计出来的回归分位系数的渐进正态性来计算置信区间。比较有代表性的是Sparsity算法，它是一种最直接且运算速度也最快的算法，但该算法得到的估计值对于随机项为独立同分布这一假设十分敏感。
3. 秩得分法。秩得分算法比较简单，但是对于大型数据处理效率较慢。
4. 重复抽样法。该方法用了MCMB算法，这种算法能够进行高效率的运算，大大节省了运算时间。重复抽样方法能够克服直接法和秩得分法的缺陷，但是对于小样本时计算出的参数估计值不够稳定。

# 问题10：非参与半参估计的意义。常用的核函数有哪些，并了解核密度估计和最优带宽的各种性质及两者之间的关联。

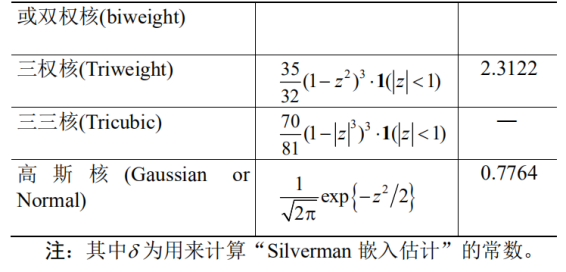
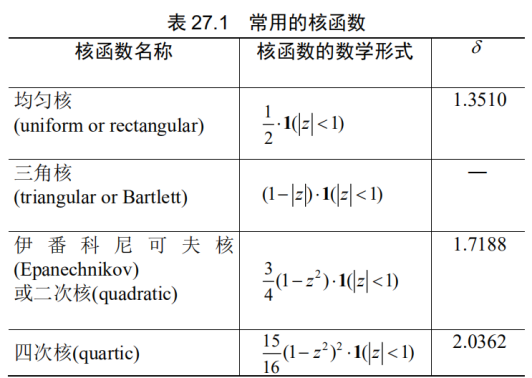
# 非参与半参估计的意义

“参数估计法”假设总体服从带未知参数的某个分布(比如正态)，或具体的回归函数，然后估计这些参数。其缺点是，对模型设定所作的假定较强，可能导致较大的设定误差，不够稳健。

“非参数估计法”一般不对模型的具体分布或函数形式作任何假定，更为稳健。缺点是要求样本容量较大，且估计量收敛的速度较慢。

作为折衷，同时包含参数部分与非参数部分的“半参数方法”，降低对样本容量的要求，又有一定稳健性。**非参及半参方法与传统的参数法互补；后者不太适用时，可考虑前者。**

# 二、常用的核函数



# 三、核密度估计的性质

由于核密度估计使用了在附近的点x来估计,而一般地，如果，则，故核密度估计通常是有偏的。

即偏差与成正比，为的同阶无穷小，记为。带宽h越大，则将使用离更远的点在估计，导致偏差增大（以的速度迅速上升）。

当，让带宽，则偏差将在大样本中消失。密度函数的二阶导数越大，即在处的曲率越大，则附近的函数值波动越大，也会引起偏差增大。

偏差还取决于核函数。

核密度估计的方差为：

故。

样本容量n越大，则方差越小；

带宽h越大，由于使用了更多观测点来估计，故方差越小。当时，让(虽然，但h趋于0的速度比样本容量的速度更慢)，则此方差将在大样本中消失。

1. **核密度估计的一致性**

当时，让带宽且，则偏差与方差在大样本下都趋于0。根据样本收敛可知，的一致估计量。

1. **核密度估计的渐近正态性**

如果核函数的条件满足，则服从渐进正态分布：据此可进行区间估计。

核密度估计的收敛速度为。由于最优带宽成正比，故这意味着非参估计量的收敛速度慢于参数估计量的通常收敛速度。

# 四、最优带宽的性质

如果带宽h越大，则附近的邻域越大，故偏差也越大(偏差与成正比)；而带宽h越大，则越光滑，即方差越小。

在选择“最优带宽”时，希望最小化均方误差(MSE)，即方差与偏差平方之和：

由于，而其中为常数，对h求导,可得一阶条件为：，。故最优带宽为。**选择最优带宽，就能保证核密度估计的一致性。**

# 问题11：了解空间计量模型，以及各个模型的优缺点、适用范围。

## 一、空间自回归模型(SAR)

也称为“空间滞后模型”(SLM)。其中，***W***为已知的空间权重矩阵（非随机），而空间依赖性仅由单一参数来刻画（待估参数大大减少）。度量空间滞后***Wy***对***y***的影响，称为“空间自回归系数”。直观来看，相邻地区的被解释变量（比如犯罪率）可能相互依赖，并最终形成一个均衡的结果。又比如，假设以地区税收为被解释变量，则不同地区的政府出于相互竞争或博弈的考虑（比如，竞相吸引FDI），在制定本地区税收时，会考虑周边地区的税收水平。

也可在方程中加入自变量：

其中，***X***为数据矩阵，包括*k*列解释变量；而为相应系数。常使用最大似然估计，推导得

由此可见，区域*j*的变量*xjr*对任意区域*i*的被解释变量都可能有影响。如果j=I，则

表明区域*i*的变量*xir*对本区域被解释变量*yi*的“直接效应”为*Sr*(***W***)*ii*，即矩阵*Sr*(***W***)主对角线上的第*i*个元素。如果将矩阵*Sr*(***W***)主对角线上的所有元素进行平均，即可得到变量*xr*的“平均直接效应”。假设所有区域的变量*xr*都变化一个单位，其对区域*i*被解释变量*yi*的“总效应”为矩阵*Sr*(***W***)的第*i*行元素之和。如果对所有区域的总效应进行平均，则得到变量*xr*的“平均总效应”。变量*xr*的“平均间接效应”为平均总效应与平均直接效应之差。

缺点：若进行OLS估计，将会导致联立方程偏差。正是由于空间依赖性，导致变量之间互相影响，从而产生内生性。

适用范围：能控制不同地区经济行为被解释变量之间的直接影响；探测由于各种空间溢出而产生的空间自相关，这种空间溢出来自于区域间存在的实质性空间相互作用，如技术扩散、要素转移、知识交流等产生的扩散和极化效应。

## 二、空间杜宾模型(SDM)

假设区域*i*的被解释变量*yi*依赖于其邻居的自变量：

其中，表示来自邻居自变量的影响，而为相应的系数向量。例如，区域i的犯罪率不仅依赖于本区域的警力，还可能依赖于相邻区域的警力。

将空间杜宾模型与空间自回归模型相结合，得到

有时也称为空间杜宾模型。

优点：方程不存在内生性，可直接进行OLS估计。

缺点：解释变量***X***与***WX***之间可能存在多重共线性。

适用范围：区域行为的空间依赖性除直接受到邻近其他地区行为的影响外，还来源于可能相互依赖的解释变量等外生变量的间接影响。

## 三、空间误差模型(SEM)

其中，扰动项***u***的生成过程为

其中，***M***为空间权重矩阵。扰动项***u***存在空间依赖性，这意味着，不包含在***X***中但对***y***有影响的遗漏变量存在空间相关性，或者不可观测的随机冲击存在空间相关性。最有效的方法为MLE估计。

优点：尽管扰动项存在自相关，但由于不存在内生性，故OLS估计是一致的。

缺点：由于忽略了扰动项的自相关而损失了效率。

适用范围：能度量因误差项之间的互动而产生的间接影响。

## 四、一般的空间计量模型

将空间自回归模型(SAR)与空间误差模型(SEM)结合起来：

其中，扰动项***u***的生成过程为

其中，***W***与***M***分别为被解释变量***y***与扰动项***u***的空间权重矩阵，二者可以相等。上面两个组合起来称为“带空间自回归误差项的空间自回归模型”(SARAR)。

也可进行MLE估计，但MLE估计量存在三个缺陷：MLE估计量可能计算不方便，特别当空间权重矩阵***W***与***M***的维度较高时；空间计量模型MLE估计量的大样本理论尚不健全；如果扰动项不服从独立同分布的正态分布，则QMLE估计量可能不一致。可以利用工具变量，通过GMM来估计SARAR模型，称为“广义空间二段最小二乘法”(GS2SLS)，计算方便且结果稳健（在异方差情况下也成立），但必须有外生变量X，故不适用于没有外生变量的纯SAR模型。

## 五、含内生解释变量的SARAR模型

其中，***Z***由内生解释变量所组成，扰动项***u***的生成过程为

## 六、空间面板模型

面板的空间自回归模型：

其中，为空间权重矩阵***W***的第*i*行，，为空间权重矩阵***W***的(*i*,*j*)元素，*ui*为区域*i*的个体效应。如果不考虑空间滞后项，则方程为标准的静态面板模型。如果*ui*与*xit*相关，则为固定效应模型；反之，为随机效应模型。

对于固定效应的空间自回归模型，可以先做组内离差变换，去掉个体效应*ui*；然后再使用类似于横截面空间自回归模型的MLE估计。对于随机效应的空间自回归模型，则可以先做广义离差变换，然后再进行MLE估计。在决定使用固定效应还是随机效应模型时，可进行通常的豪斯曼检验。

# 问题12：如何使用多项Logit回归，混合Logit型，嵌套Logit模型对数据进行处理，并对结果进行分析。

（结果分析在问题13）

## 一、多项Logit回归

使用随机效用法，假设个体*i*选择方案*j*所能带来的随机效用为

其中，解释变量只随个体*i*而变，不随方案*j*而变。比如，个体的性别、年龄、收入等特征。这种解释变量被称为“只随个体而变”或“不随方案而变”。系数表明对随机效用的作用取决于方案*j*。当且仅当方案*j*带来的效用高于所有其他方案时，个体*i*选择方案*j*。

个体i选择方案j的概率为

其中，j=1所对应的方案为参照方案。可用MLE估计多项Logit模型。个体i的似然函数为

其对数似然函数为

其中，**1**(·)为示性函数，即如果括号中的表达式成立，则取值为1；反之，取值为0。将所有个体的对数似然函数加总，即得到整个样本的对数似然函数，将其最大化则得到系数估计值。

## 二、混合Logit模型

假设个体i选择方案j所能带来的随机效用为

其中，解释变量既随个体*i*而变，也随方案*j*而变；而解释变量*zi*只随个体*i*而变。经过类似的推导可知，个体*i*选择方案*j*的概率为

需要选择一个参照方案，然后令。

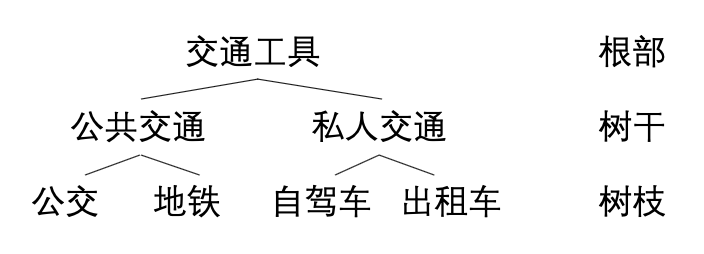
当各方案本身的特征不重要，或缺乏有关方案特征的数据时，常使用多项Logit模型。如果需要考虑不同方案的特征，则应使用条件Logit模型或混合Logit模型。另外，在这些多值选择模型中，由于被解释变量的分布必然为“多项分布”，故一般不必使用稳健标准误，使用普通标准误即可。如果数据为聚类样本，则仍应使用聚类稳健的标准误。

在多项Logit与混合Logit模型中，对参数估计值的解释是以参照方案为转移的。将多值选择模型中的任何两个方案单独挑出来，都是二值Logit模型，此假定称为“无关方案的独立性”（IIA），多项Logit和条件Logit模型都服从IIA假定。在实践中，如果不同方案之间很类似（红车蓝车问题），则IIA假设不一定满足，这是多项Logit、条件Logit与混合Logit模型的共同缺点。

对于IIA问题，检验方法之一为豪斯曼检验。其基本思想是，如果IIA假定成立，则去掉某个方案不影响对其他方案参数的一致估计，只是降低了效率。这意味着，在IIA原假设成立的情况下，去掉某个方案后子样本的系数估计值与全样本的系数估计值没有系统差别。检验方法之二为Small-Hsiao检验，但这两个检验的小样本性质都不好，故其结论只具有参考价值。

## 三、嵌套Logit模型

多项Logit、条件Logit与混合Logit都须满足IIA假定，但在实践中，如果方案之间比较类似，则IIA假设可能不满足。解决方法之一是，把相似的方案归入一组，允许同组内的方案相关，但不同组的方案相互独立。使用嵌套Logit的前提是，各备选方案之间应有一个清晰的嵌套结构。



假设有*J*个树干，树干*j*包含*Kj*个树枝。个体*i*选择树干*j*树枝*k*方案所能带来的效用为。其中，解释变量*zj*只随树干而不随树枝方案而变（故其系数可随*j*而变），而*xjk*同时随树干与树枝方案而变（故其系数没有下标，不随*j*、*k*而变）。

假定扰动项服从“广义极值分布”(GEV)，其累积分布函数为

其中，函数G(·)的形式为

参数与的相关系数呈反向变动的关系，故称为“不相似参数”。如果，则所有方案的扰动项均不相关，满足IIA假定，回到多项Logit、条件Logit或混合Logit的情形。选择树干*j*树枝*k*方案的概率为

其中，*pj*为选择树干*j*的概率，*pk|j*为在选择树干*j*的情况下，选择树枝*k*的条件概率；而*Ij*的定义为

*Ij*称为“包含价值”或“对数和”。写出整个样本的对数似然函数，称为“全信息最大似然估计”(LIML)。进行FIML估计后，可对联合假设“”进行似然比检验。如果接受原假设，则IIA假定成立，不需要使用嵌套Logit，可直接使用多项Logit、条件Logit或混合Logit模型。另外，也可以考虑更多层的嵌套结构。

以上假设所有解释变量都随方案而变。如果解释变量包括不随方案而变的个体特征，则个体*i*选择树干*j*树枝*k*方案所能带来的效用为。

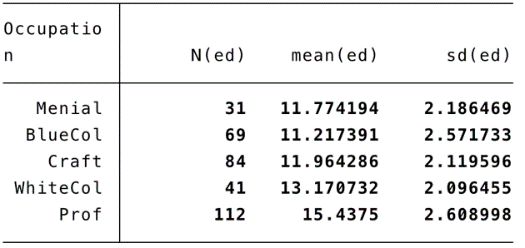
# 问题13：实例分析题来自第十二章

## 例1：多项Logit、多项Probit

问卷调查将受访者职业分为五类(occ)，即服务人员(menial)、蓝领(blue collar)、工匠(craft)、白领(white collar)以及专业人士(professional)。解释变量包括：是否白人(white)、受教育年限(ed)以及工龄(exper)。这三个解释变量都只依赖于个体，而不依赖于方案，故使用多项Logit或多项Probit回归。

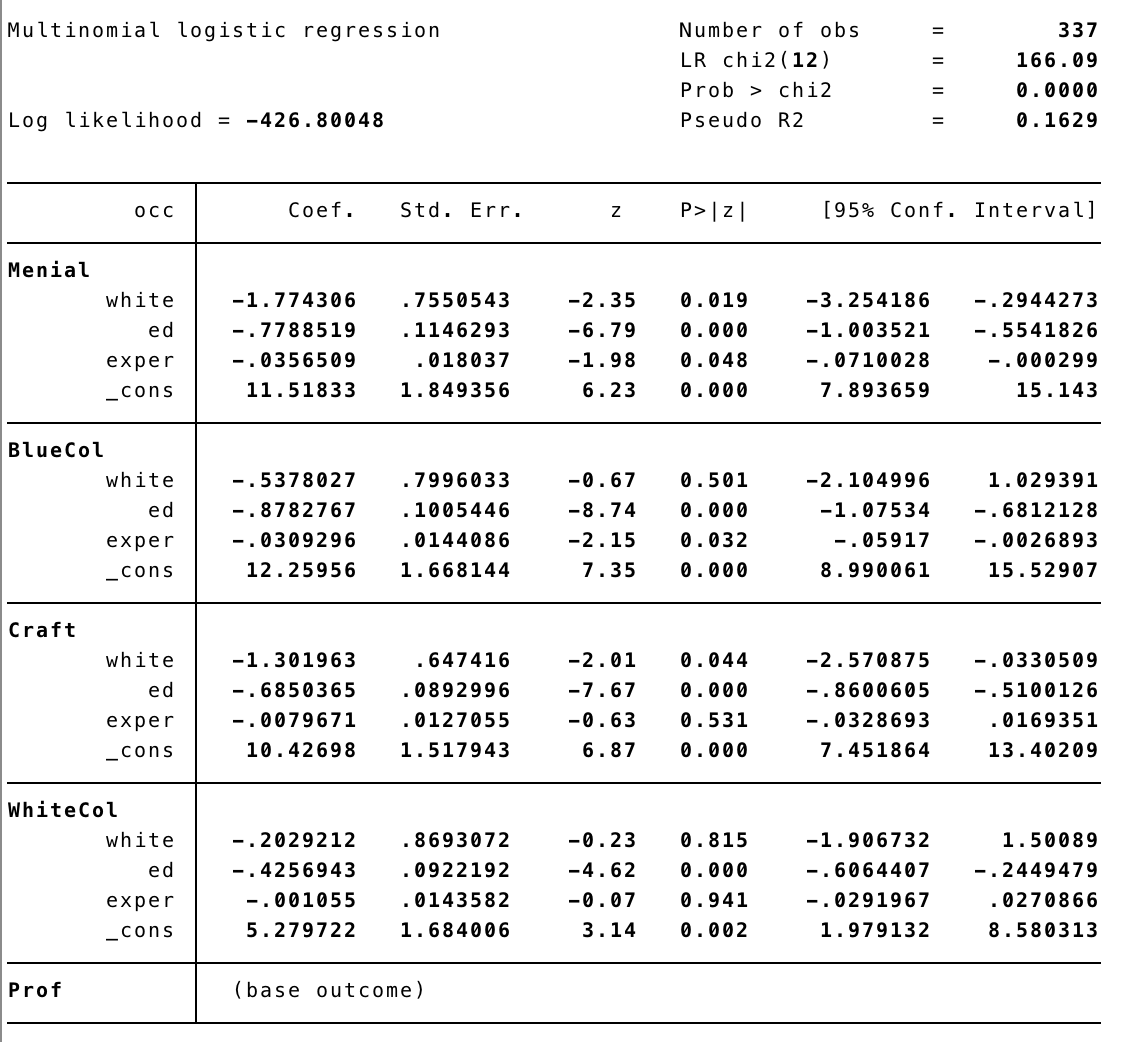
### 1.考察受教育年限与职业的关系

罗列受教育年限的样本容量、均值与标准差（按职业分为5个子样本）



在样本中，专业人士最多，且平均受教育年限最高；而蓝领的受教育程度最低。

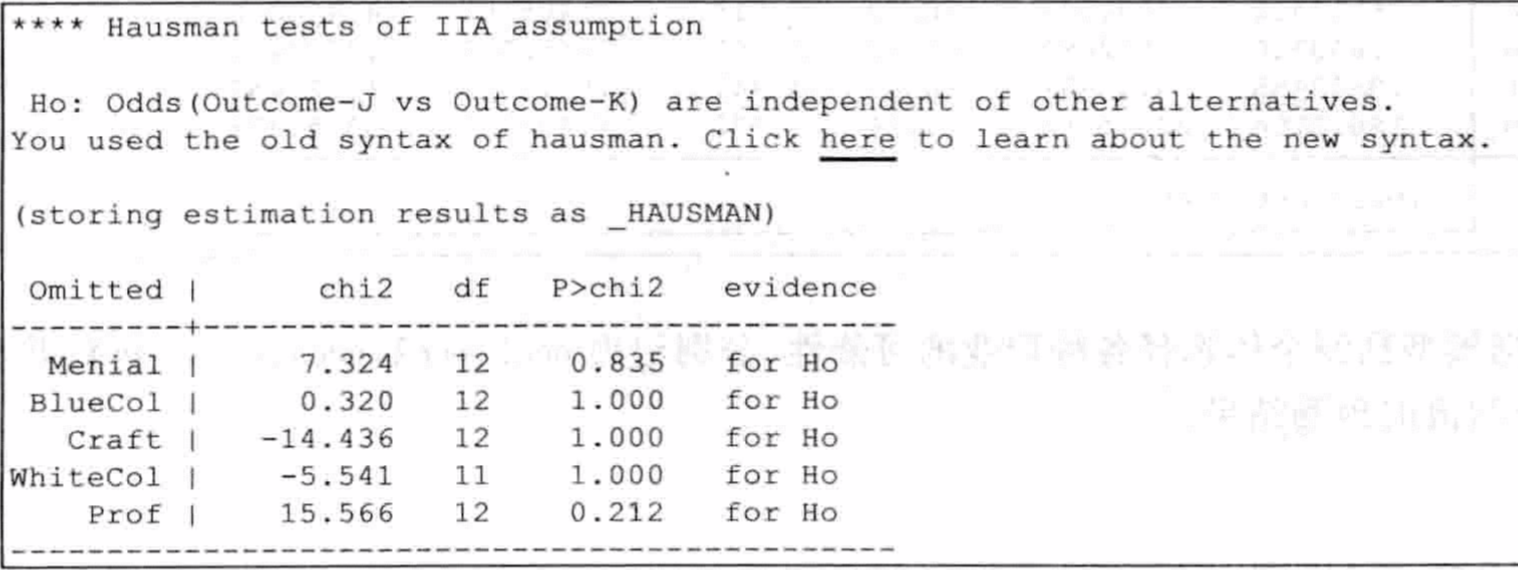
### 2.进行多项Logit回归



由于没有指定参照方案，故命令自动选择观测值最多的方案（专业人士）为参照方案。上表显示，在5%的显著性水平上，给定其他变量，白人更不可能选择服务业或工匠（menial和craft中的white的p值都小于0.05）；但是否白人对于选择蓝领或白领没有显著影响（bluecol和whitecol中的white的p值都大于0.05）。受教育程度越高，越不可能选择除专业人士以外的职业（表中4种职业的ed的p值都小于0.05）。工龄越长，越不可能选择服务业或蓝领（menial和bluecol中的exper的p值都小于0.05）；工龄对于选择工匠或白领无显著影响（craft和whitecol中的exper的p值都大于0.05）。

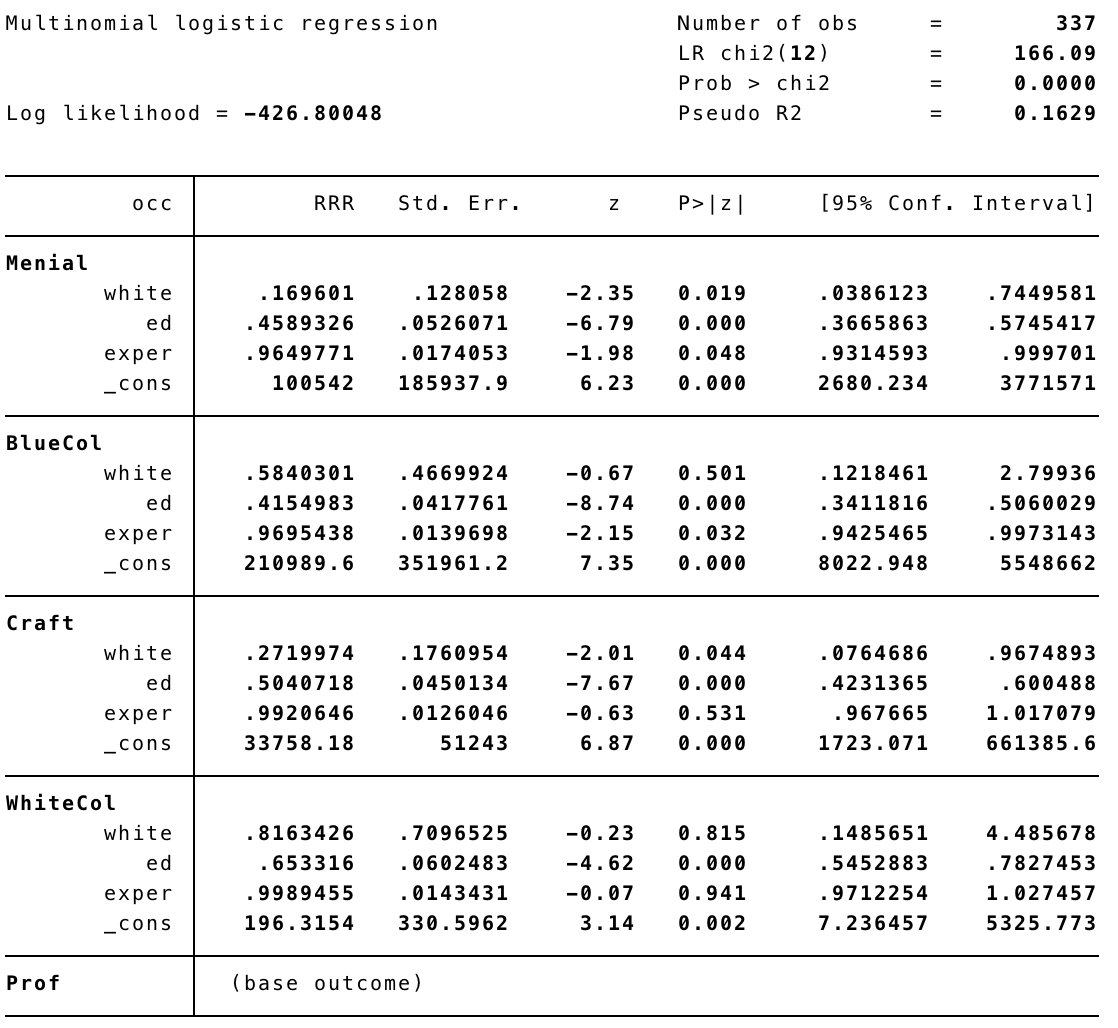
### 3.豪斯曼检验（检验IIA假定是否满足）

因为IIA假定是多项Logit模型的前提

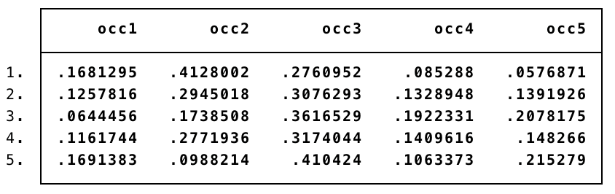


M、B、C、W这四行豪斯曼检验结果显示，去掉四个非参照方案中的任何一个方案，都不会拒绝IIA的原假设。由于使用了选择项“base”，故最后一行的计算去掉参照方案(Prof)，而以剩余方案中观测值最多的方案作为参照方案的检验结果，同样也不拒绝IIA假设（五行的p值都大于0.05）。

### 4.显示相对风险比率



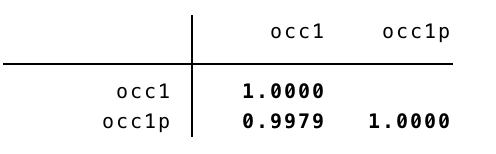
### 5.预测个体选择各种职业的可能性



第一个个体最有可能选择第二种职业……

**6.**选择其他职业作为参照方案，系数估计值发生变化。多项Probit和多项Logit的系数估计值不同，但二者的系数不具可比性；具有可比性的是两个模型的预测概率。

**7.计算多项Probit模型所预测**的各种职业的选择概率，计算两个模型职业预测概率的相关性。



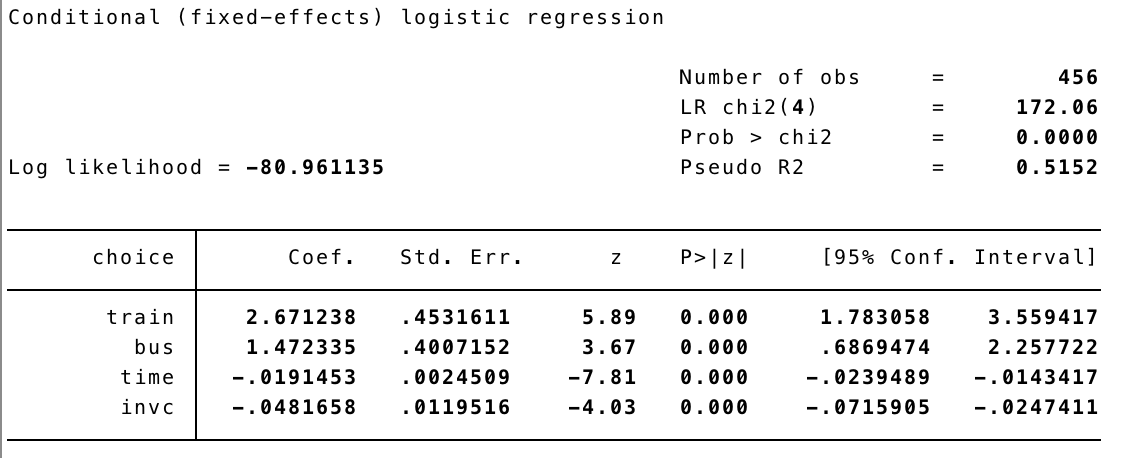
两个模型所预测的职业选择概率高度一致，相关系数均在99%以上。

## 例2：条件Logit、混合Logit

该数据集的观测单位为152组人群（每组可视为一个小型旅行团），每个人群从以下三种度假旅行方式中选择一种，即火车、长途大巴或自驾车。随方案而变的解释变量包括time(总旅行时间)与invc(乘车成本)；不随方案而变的解释变量包括hinc(家庭收入)与psize(旅行团人数)。进行条件Logit与混合Logit估计。

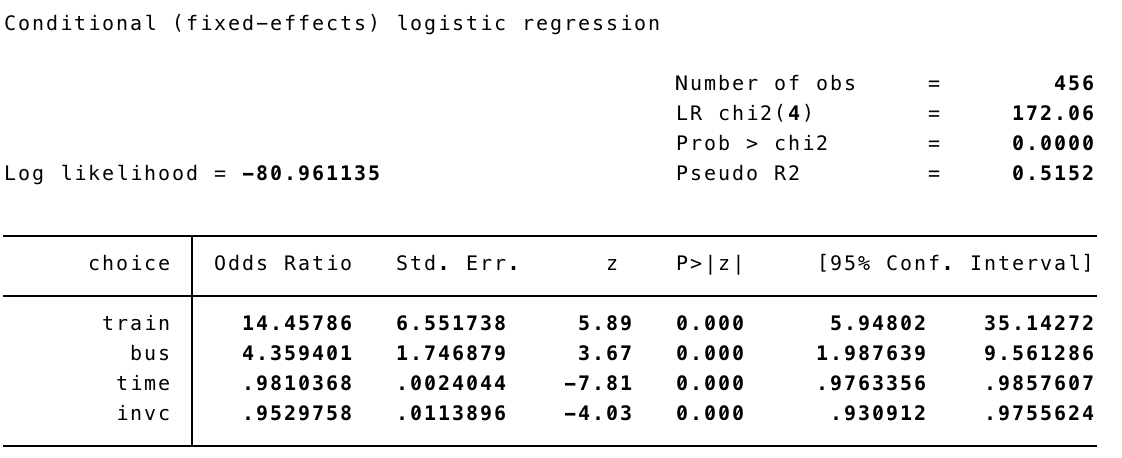
### 1.条件Logit估计

使用随方案而变的解释变量：总旅行时间、乘车成本、虚拟变量train、虚拟变量bus（以旅行方式car为参照方案）。



上表显示，如果其他解释变量(time,invc)的取值相同，则旅行团最有可能选择火车，其次为长途大巴（按系数大小？）。另外，一个方案的总旅行时间越长，乘车成本越高，则选择该方案的概率越低（train、bus与time、invc的系数正负号相反？）。然而，由于这是非线性模型，故不易通过系数估计值来评价边际效应。

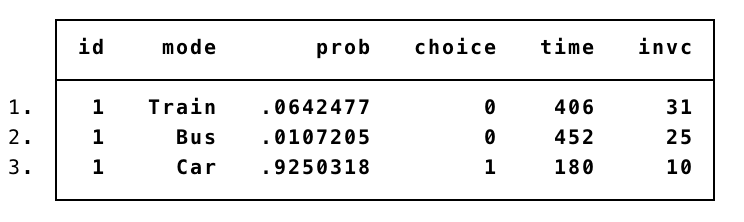
### 2.计算风险比率



变量time的风险比率为0.98，这意味着在给定其他变量的情况下，一个方案的总旅行时间每增加1分钟，则选择此方案的概率将乘以0.98，即下降2%；invc解释同上。虚拟变量bus的风险比率为4.36，这意味着，如果各方案的时间与成本均相等，则旅行团选择长途大巴的概率是选择自驾车（参照方案）概率的4.36倍；train解释同上。

### 3.计算条件Logit模型的预测概率

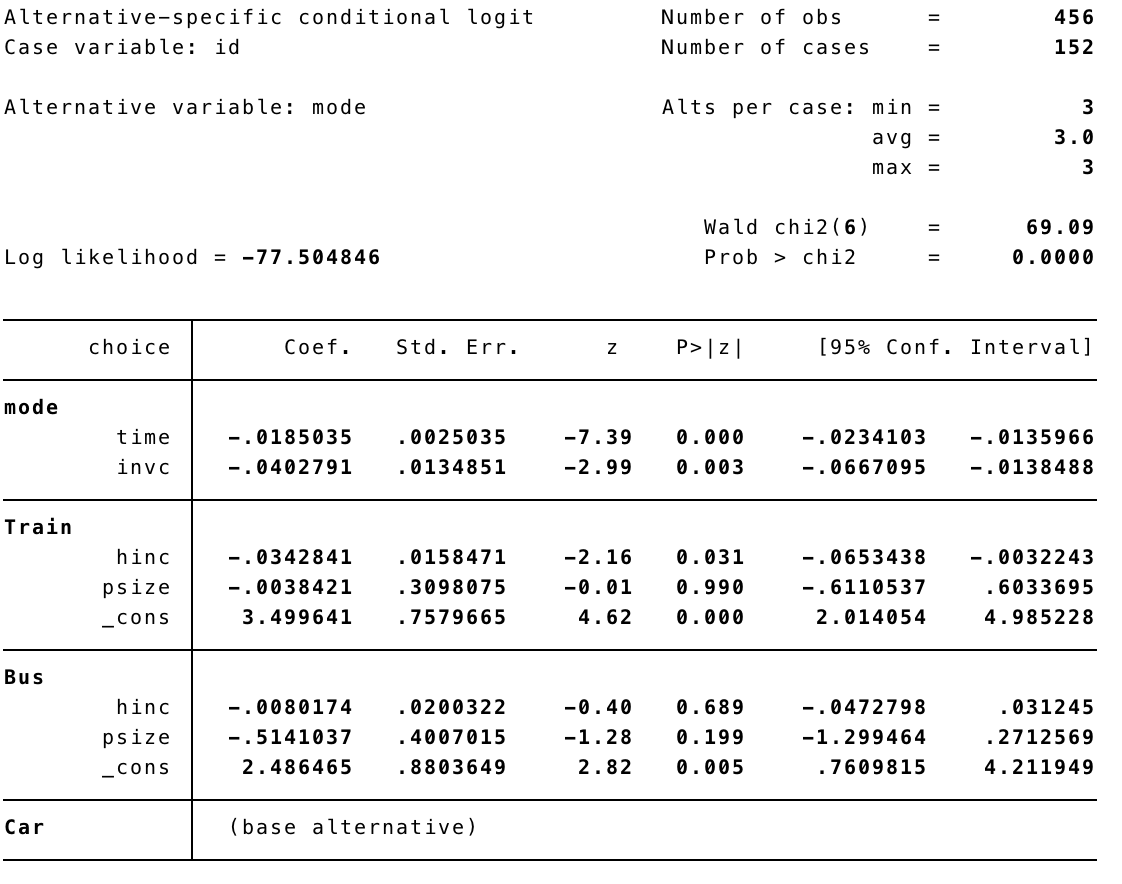
第一个旅行团的预测概率：



第一个旅行团实际选择car，且模型预测选择car的概率高达0.925，正好是旅行时间与成本最低的方案。

### 4.混合Logit模型估计

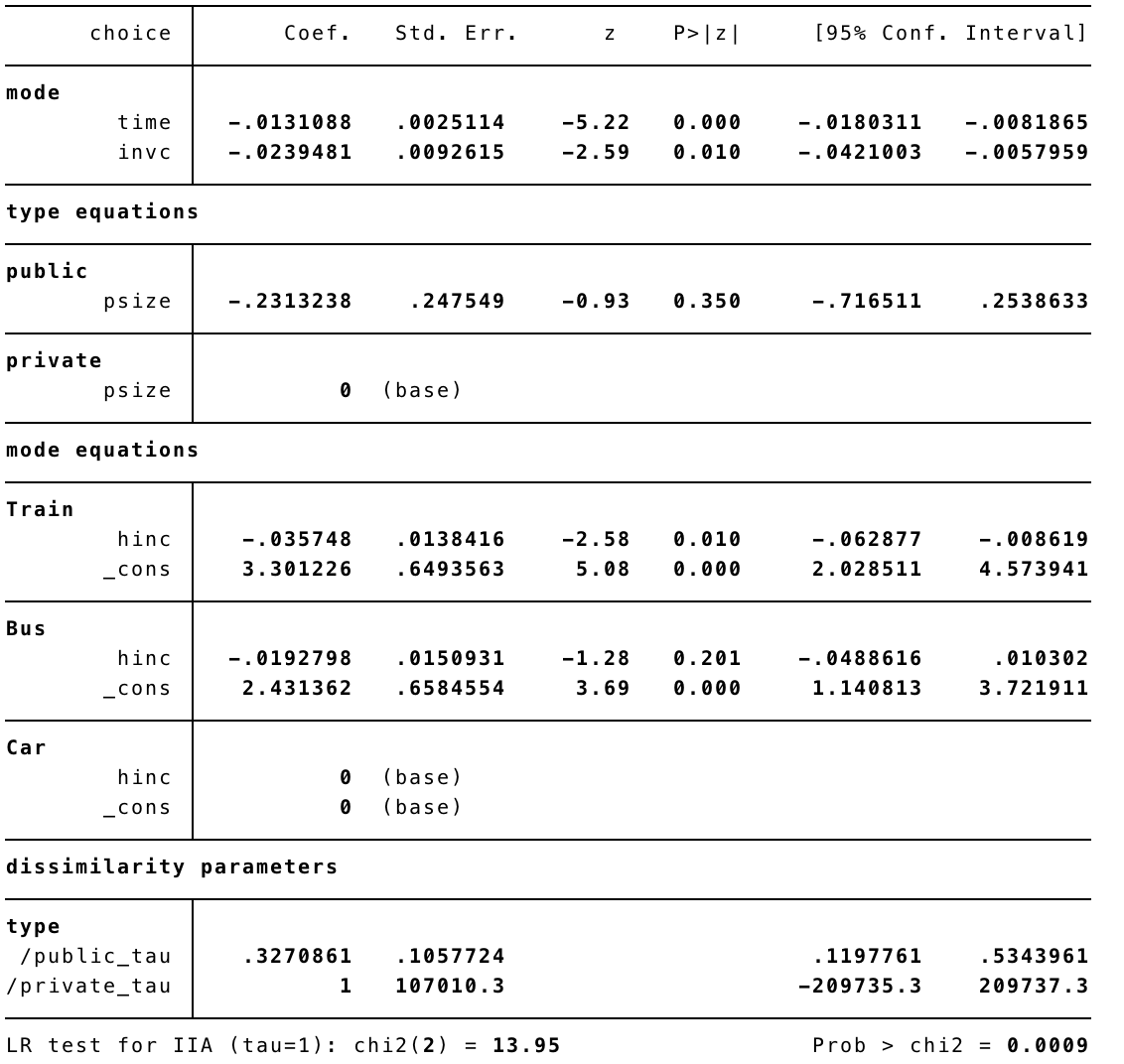
将家庭收入与旅行团人数也包括进来



上表显示，家庭收入越高，越不倾向于选择火车(train-hinc的系数显著为负)；但对于选择长途大巴则无显著影响(bus-hinc的系数不显著)。旅行团规模没有显著影响(psize的系数都不显著)。旅行时间与乘车成本的估计系数依然显著为负，且与前面条件Logit模型的估计系数很接近。

## 例3：嵌套Logit

将例2中的train和bus归入一组（称为public），而car自成一组（称为private），将定义嵌套结构的变量记为type。



最后一行的似然比检验强烈拒绝IIA假定（p值小于0.05），故应使用嵌套Logit模型。旅行团规模(psize)对于选择公共交通(public)或私人交通(private)没有显著影响。家庭收入(hinc)越高，越不倾向于选择火车，但对选择长途大巴无显著影响。变量time与invc的系数依然显著地为负，但估计值与前面的条件Logit或混合Logit模型的估计值有一定差距（由于后者没有考虑Train与Bus两个公共交通方案间的相关性，故可能高估了变量time与invc的作用）。public组的不相似参数为0.327；而对于private组，由于只有一个方案，故不相似参数标准化为1。

**异方差** P2

**Logit模型** P6

（1）代表存活的概率是死亡概率的3倍。

（2）意味着*xj*增加一单位引起几率比增加12%。

**对空间计量模型的认识、各个模型的适用范围** P11

**面板数据** P8

**优点**

1、可以解决遗漏变量问题。遗漏变量虽然可以用工具变量法解决，但有效的工具变量常常难以寻找。遗漏变量常常是由于不可观测的个体差异或异质性造成的，如果这种个体差异“不随时间而改变”，而面板数据提供了解决遗漏变量问题的又一利器。

2、提供更多个体动态行为的信息：由于面板数据同时有横截面与时间两个维度，有时它可以解决单独的截面数据或时间序列数据不能解决的问题。

3、样本容量较大：由于同时有截面维度与时间维度，通常面板数据的样本容量更大，从而可以提高估计的精度。

证明 T=2时yi1=xi1’β+zi’δ+ui+εi1 yi2=xi2’β+zi’δ+ui+εi2 FD: yi2- yi1=β(xi2- xi1)’+(εi2-εi1) FE: yit- 所以FD估计和FE估计量的模型表示形式是一致的。

**变系数**：系数模型的两大类 P9

**分析题**（1）上表显示，在5%的显著性水平上，给定其他变量，白人更不可能选择服务业或工匠，但是是否白人对于选择蓝领还是白领没有显著影响。受教育程度越高，越不可能选择除专业人士以外的职业。工龄越长，越不可能选择服务业或蓝领；工龄对于选择工匠或白领没有显著影响。（2）从上表可知，变量time的风险比率为0.98，这意味着在给定其他变量的情况下，一个方案的总旅行时间每增加1分钟，则选择该方案的概率将乘以0.98，即下降2%。变量inve的风险比率可类似地解释。另一方面，虚拟变量bus的风险比率为4.36，这意味着，如果各方案的时间与成本均相等，则旅行团选择长途大巴的概率是选择自驾车概率的4.36倍；虚拟变量train的风险比率也可类似地解释。（3）上表最后一行的似然比检验强烈拒绝IIA假定，故应使用嵌套Logit模型。另外，旅行团规模对于选择公共交通或私人交通没有显著影响。家庭收入越高，越不倾向于选择火车，但对选择长途大巴无显著影响。变量time与inve的系数依然显著的为负，但估计值与前面的条件Logit或混合Logit模型的估计值有一定差距。最后，上表下部还显示，public组的不相似参数为0.327；而对于private组，由于只有一个方案，故不相似参数标准化为1。

**自相关性** P2

**DW检验方法的论述及缺陷**：P2

**工具变量能解决什么问题**： P5

模型中一个或多个解释变量与扰动项相关的问题，即内生性问题。

**条件**：（1）相关性：工具变量与内生解释变量相关，即Cov(xt,pt)≠0。(2)外生性：工具变量与扰动项不相关，即Cov(xt,ut)=0。

**究竟该使用OLS还是工具变量法呢？** P6

假设存在方差外的工具变量。如果所有解释变量都是外生变量，则OLS比工具变量法更有效。在这种情况下使用工具变量法，虽然估计量仍然是一致的，但相当于无病用药，反而增大估计量的方差。反之，如果存在内生解释变量，则OLS是不一致的，而工具变量法是一致的。

**为什么需要非参数与半参数估计。** P10

（1）参数估计法的缺点：对模型设定所作的假定较强，可能导致较大的设定误差。比如，如果真实总体并非正态。甚至偏离正态较远，则在正态分布前提下所作的统计推断可能有较大偏差。换言之，由于参数估计法对模型设定的依赖性较强，故可能不够稳健。

（2）非参数估计法一般不对模型的具体分布作任何假定，更为稳健，但缺点是，要求的样本容量较大，而估计量收敛到真实值的速度也较慢。作为一种折中，同时包含参数部分和非参数部分的半参数方法应运而生，它降低了对样本容量的要求，又具有一定稳健性。

**断尾回归为什么不能使用最小二乘估计？** P7

**断尾回归的极大似然估计有什么可以注意的地方？**

今年新增问题：分位数回归与OLS和MLE的关联和区别？（需自己补充）

假设不知道残差的分布，但残差满足期望为0，且方差有限。1. 新来一个样本，在预测意义下，ols只能提供Y条件期望的预测，分位数回归还可以给出Y的各个分位数的预测，能提供给多的信息。2. 和线性回归相比，分位数回归对异常值更加稳健。3. 在用ols估计后，把组内预测误差作为残差分布的估计，这在残差分布iid时侯是可行的做法，但在异方差时候无法这样做。分位数回归则可以处理某些异方差模型。