- 自相关的概念，产生原因，如何检验，处理的方法，产生的影响。

概念：对于 K 元线性回归模型：$y_i = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + \varepsilon$，如果不同随机误差项之间存在相关关系，即 cov(,,)!=0,则称模型存在自相关性

原因：

1. 模型中遗漏了重要的解释变量(消费，收入；消费习惯，各期相关)

2. 经济变量的惯性作用(前后期之间互相关联)

3. 模型设定ε不当的影响

4. 一些随机干扰因素的影响

检验：

主要相关性检验有四种：图示法、回归检验法、杜宾-瓦森检验法（D.W.）、拉格朗日检验（GB）。

最好的检验方法应该是 GB 检验，适用于高阶序列相关及模型中存在滞后变量的情形。D.W.检验中，存在一个不能确定的 D.W.值区域，且仅能检测一阶自相关，对存在置后被解释变量的模型无法检验

处理：

1. 使用 ols+异方差自相关稳健的标准误，只改变标准误的估计值，不改变回归系数的估计值

2. 使用 ols+聚类稳健的标准误(面板数据)

3. 可行广义最小二乘法

4. 修改模型设定

影响：

1. 自相关不影响 OLS 估计量的线性和无偏性，但使之失去有效性

2. 自相关的系数估计量将有相当大的方差

3. 自相关系数的 T 检验不显著

4. 模型的预测功能失效

- 异方差的概念，产生原因，如何检验，处理的方法，产生的影响。

概念：对于模型…如果出现$Var(\varepsilon)=\sigma_i^2$ ，即对于不同样本点，随机误差项的方差不再是常数，而互不相同，则认为模型出现了异方差

原因：

1. 模型中遗漏了某些解释变量

2. 模型函数形式的设定误差

3. 样本数据的测量误差

4. 随机因素的影响

检验： 图示检验法、Goldfeld - Quandt 检验法、White 检验法、Park 检验法,Gleiser 检验法。

处理：

1. 使用 ols+稳健标准误

2. 广义最小二乘法

3. 加权最小二乘法

4. 可行广义最小二乘法

影响：如果线性回归模型的随机误差项存在异方差性，会对模型参数估计、模型检验及模型应用带来重大影响

1）不影响模型参数最小二乘估计值的无偏性

2）参数的最小二乘估计量不是一个有效的估计量

3）对模型参数估计值的显著性检验失效

4）模型估计式的代表性降低，预测精度精度降低。

- 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。

遗漏变量:遗漏的解释变量，与解释变量相关或者不相关，不相关不用考虑，不会影响。

1. 增加控制变量
2. 随机试验和自然实验(准实验法)
3. 工具变量法
4. 面板数据

概念：多重共线性是指线性回归模型中的解释变量之间由于存在精确相关关系或高度相关关系而使模型估计失真或难以估计准确。

原因：

1. 经济变量相关的共同趋势
2. 滞后变量的引入
3. 样本资料的限制

检验：

1. 相关系数检验
2. 辅助回归模型检验
3. 方差膨胀因子法
4. 直观判断法

处理：

1. 排除引起共线性的变量影响
2. 差分法
3. 减小参数估计量的方差岭回归法

4. 简单相关系数检验法

影响：

1. 参数估计量经济含义不合理

2. 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

3. 模型的预测功能失效

虚拟概念：虚拟变量又称虚设变量、名义变量或哑变量，用以反映质的属性的一个人工变量，是量化了的自变量，通常取值为 0 或 1。

作用：

1. 分离异常因素的影响

2. 检验不同属性类型对因变量的作用

3. 提高模型的精度， 扩大了样本容量（增加了误差自由度，从而降低了误差方差）

- 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。

概念：如能将内生变量分成两部分，一部分与扰动项相关，另一部分与扰动项不相关，可用与扰动项不相关的那部分得到一致估计。这种分离常借助另一“工具变量”来实现。

能够解决解释变量和扰动项相关的回归估计问题从而得到一致估计。

需要满足：

1. 相关性：工具变量与内生解释变量相关，即 $Cov(z_t, p_t) \neq 0$ 。

2. 外生性：工具变量与扰动项不相关，即 $Cov(z_t, p_t) = 0$ 。

检验：

1. 弱工具变量检验(包含很少与 x 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，即样本容量很小也很难收敛到真实的参数值

2. 过度识别检验

3. 究竟用 ols 还是工具变量法:对解释变量内生性的检验(豪斯曼)

4. 不可识别检验(检验秩条件是否成立)

- 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。

如果被解释变量为离散变量或者虚拟变量时，使用离散选择模型(二值或者多值)，也就是面板二值选择模型。

以二值选择（被解释变量取值为 0 或 1）为例，当被解释变量取 1 的概率为标准正态分布时，使用 probit 模型；当被解释变量取 1 的概率为 logistic 分布时，使用 logit 模型。

当模型的被解释变量为二值变量时，线性回归方法一般不再适用，需要采用其他方法。二值选择模型关注的是自变量的变动对因变量取值的概率的影响。

实际应用中，常用 Probit 模型和 Logit 模型对二值选择模型进行估计，两者的区别在于对连接函数具体形式的设定不同。多值:多项 p 和多项 l 模型，条件和混合 l 模型。应用:比如是否吸烟，喝酒对肠癌等病的影响，以及某些二分类因素对某疾病的影响等

- 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。

概念:对于线性模型 $y_i = x_i^{'}\beta + \varepsilon_i$，假设由于某种原因，只有满足 $y_i \geq c$ 的数据才能观测到，因此，当 $y_i < c$ 时，没有任何有关 $\{x_i, y_i\}$ 的数据，例如企业销售收入，统计局只统计>=1000 的情况，这样被假释变量则在左边断尾。

OLs 回归中，扰动项和解释变量相关，导致不一致的估计，收敛不到真实的系数

值，同时可能出现预测值<=c(断点处)的情形；mle 可以得到一致估计，由于随机变量断尾后，其概率分布也会随着发生变化，OLS 回归会导致不一致的结果，所以一般用 MLE 进行估计。极大似然估计需要建立似然函数，然后取对数求导，解似然方程。被解释变量受限的另一种情况是对于线性模型，当 $y_i \geq c$ 时，所有 y_i 都被归于 c，这种数据为归并数据，虽然有所有数据，但是时一个离散点和连续分布，Ols 是不一致的估计，tobit 就是 mle 估计，写出整个样本的似然函数，然后 mle 估计。

- 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。在处理面板数据时，应该使用固定效用还是随机效用模型。

概念：面板数据，也叫“平行数据”，是指在时间序列上取多个截面，在这些截面上同时选取样本观测值所构成的样本数据。或者说他是一个 $m \times n$ 的数据矩阵，记载的是 n 个时间节点上，m 个对象的某一数据指标。

分类：

1. 混合估计模型(在时间和截面上都不存在显著性差异，就可以把面板数据混合在一起用最小二乘法)
2. 固定效应模型(对不同截面或者时间序列，模型的截距是不同的。个体，时点，个体时点)
3. 随机效应模型(a 是与 x_i^t 不相关的随机变量，不同个体有不同 a)

其它：

动态，系变数，面板数据的向量自回归，非均衡面板数据，离散

优缺点：

第一， 面板数据可以解决遗漏变量问题。遗漏变量通常由于不可观察的个体差

异或“异质性”导致，如果这一异质性不随时间变化，那么面板数据便提供了解决遗漏变量的利器，而这是截面数据不能解决的。

第二， 面板数据提供了个体的动态行为信息。面板数据兼具横截面和时间两个维度，可以解决横截面数据和面板数据单独不能解决的问题。

第三， 两个维度的数据使得面板样本容量大幅增加，与横截面相比，明显提高估计的精确度，因为很多估计量和检验都是在大样本下得到的渐进分布。

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，收集成本较高，不易获得。原假设u_i与x_i^t、z_i不相关(即r_e为正确模型)，成立r_e，f_e都成立，但r_e更有效，不成立则f_e成立（豪斯曼，不适用异方差）

- 长面板数据的特点，和短面板数据的差异。对长面板数据，样本容量较大时，可建立什么模型，简要叙述该模型。

特点：长面板数据时间维度 T 较大，信息较多，可以放宽扰动项独立同分布的假定。短面板时间维度较小，每个个体的信息较少，无法探讨扰动项是否存在自相关，故假定其独立同分布。

大，信息较多，故可以放松这个假定，考虑$\{\varepsilon_{it}\}$可能存在的异方差与自相关。在长面板中，由于n相对于T较小，对于可能存在的固定效应，只要加入个体虚拟变量即可（即 LSDV 法）。对于时间效应，可以通过加上时间趋势项或其平方项来控制（由于T较大，如果加上时间虚拟变量，则可能损失较多的自由度）。

为此，考虑以下模型：

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} \tag{16.1}$$

其中，$\boldsymbol{x}_{it}$可以包括常数项、时间趋势项（或其平方项）、个体虚拟变量以及不随时间变化的解释变量z_i。下面考虑扰动项$\{\varepsilon_{it}\}$存在异方差或自相关的几种情形。

（1）记个体i的扰动项方差为$\sigma_i^2 \equiv \mathrm{Var}(\varepsilon_{it})$。如果存在$\sigma_i^2 \neq \sigma_j^2 (i \neq j)$，则称扰动项$\{\varepsilon_{it}\}$存在“组间异方差”（groupwise heteroskedasticity）。

（2）如果存在$\mathrm{Cov}(\varepsilon_{it}, \varepsilon_{is}) \neq 0 (t \neq s, \forall i)$，则称扰动项$\{\varepsilon_{it}\}$存在“组内自相关”（autocorrelation within panel）。

（3）如果存在$\mathrm{Cov}(\varepsilon_{it}, \varepsilon_{jt}) \neq 0 (i \neq j, \forall t)$，则称扰动项$\{\varepsilon_{it}\}$存在“组间同期相关”（contemporaneous correlation）或“截面相关”（cross-sectional correlation）。比如，对于省际面板数据（provincial panel data），相邻省份之间的同期经济活动可能通过贸易或投资相互影响。这种相关也称为“空间相关”（spatial correlation）。

对于扰动项$\{\varepsilon_{it}\}$可能存在的组间异方差、组内自相关或组间同期相关，主要有两类处理方法。方法一，继续使用 OLS（即 LSDV）来估计系数，只对标准误差进行校正（即下文的面板校正标准误差）。方法二，对异方差或自相关的具体形式进行假设，然后使用可行广义最小二乘法（FGLS）进行估计。

- 为什么需要分位数回归。分位数回归的估计方法是什么。

之前的回归模型中，考察x对y的条件期望$E(y|x)$的影响，是均值回归，关心的是x对整个条件分布$y|x$的影响，而$E(y|x)$只是刻画$y|x$集中趋势的一个指标。如$y|x$不是对称分布，则$E(y|x)$很难反映整个条件分布全貌。

为此，Koenker and Bassett（1978）提出“分位数回归”（Quantile Regression，简记 QR），使用残差绝对值的加权平均（比如，$\sum_{i=1}^{n}|e_i|$）作为最小化的目标函数，故不易受极端值影响，较为稳健。更重要的是，分位数回归还能提供关于条件分布$y|\boldsymbol{x}$的全面信息。下面首先回顾有关总体

方法：

1）点估计：单纯形算法（参数有较好稳定性，大型数据速度慢），内点算法（对大量观察值和少量变量的数据集运算效率高），平滑算法（适合处理有大量观察值和很多变量的数据集）

2）区间估计：直接估计（根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），秩得分法（简单，对大数据处理慢），重复抽样（使用 MCMB，能进行高效率运算，克服直接发和秩得分法的缺陷，小样本计算的参数估计值不稳定）

- 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。

意义：参数估计法，即假设总体服从带未知参数的某个具体分布(比如正态分布)，然后估计这些参数。但其对模型设定所作的假设较强，可能导致较大的‘设定误差’，比如，真实总体并非正太，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，而非参一般不对模型的具体分布做任何假定，故更为稳健，但其要求样本容量较大，而估计值收敛到真实值速度也较慢，半参则是一种折中，他降低了对样本容量的要求，又具有一定的稳健性，非参和半参与传统参数估计方法是互补关系，当后者不适用时，则考虑前者。

核函数：均匀核，三角核，二次核，四次核，三权核，三三核，高斯核。核密度估计性质：一致性，渐近正态性

联系：带宽的大小决定了核密度估计函数（KDE）的平滑（smooth）程度，带宽越小越 undersmooth，带宽越大越 oversmooth 果带宽不是固定的，而是根据样本的位置而变化(其变化取决于估计的位置（balloon estimator）或样本点（逐点估计 pointwise estimator）)，则会产生一种特别有力的方法，称为自适应或可变带宽的核密度估计。自适应带宽的核密度估计方法是在固定带宽核密度函数的基础上，通过修正带宽参数为而得到的.

- 了解空间计量模型，以及各个模型的优缺点、适用范围。

空间自回归模型(滞后，最大似然估计）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。在经典线 性回归模型中引入空间滞后因变量：

$$Y = \rho WY + X\beta + \varepsilon, \varepsilon \sim N(0, \sigma^2 I)$$

空间误差模型：空间自相关在传统模型中往往被认为是噪音，它实际上 度量了邻近单元因变量的误差冲击对本单元观测值的影响 程度。空间误差模型中，空间自相关反映在误差项中：

$$Y = \boldsymbol{X}\beta + \mu,\ \mu = \lambda W\mu + \varepsilon,\ \varepsilon \sim N\left(0, \sigma^2 \boldsymbol{I}\right)$$

一般的空间计量模型(自回归和误差结合起来)：

❖空间滞后模型与空间误差模型的选择

统计依据：拉格朗日乘数检验统计量

- LM-Lag & Robust LM-Lag 对是否需要引入空间滞后项进行检验（显著为是，不显著为否）
- LM-Error & Robust LM-Error 对是否需要引入空间误差项进行检验（显著为是，不显著为否）
- 若两个稳健检验均显著，而稳健 LM-Lag 的 p 值小于稳健 LM-Error 的 p 值，则在模型中引入空间滞后项。

理论依据：经济增长中跨区域外溢产生于技术扩散和金融外部性，随机冲击在区域之间的转移只起到了很小的作用，因此，空间滞后模型更容易从经济意义上进行解释。

顾名思义，空间误差模型说明空间影响是在误差中，也就是说空间权重矩阵放在无法检测到的误差项中；空间之后模型则表明空间关联是在时间中，空间权重矩阵出现在以前的相关变量中。

空间杜宾模型：除了邻近地区的空间溢出效应外，如果在空间上邻近区域的解释变量对区域行为也有影响，需要采用空间杜宾 模型（SDM）：

$$Y = \rho WY + X\beta + \gamma WX + \varepsilon$$

SDM 模型反映了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。

地理加权回归：试图利用空间的非稳态性，使得变量间的关系随着空间的变化而变化；模型通过对回归方程残差项加权体现其空间变化。

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

$$U_{ij} = x_i'\beta_j + \varepsilon_{ij}, (i = 1,2,\cdots,n; j = 1,2,\cdots,J)$$

解释变量 x_i 只随个体 i 而变，不随方案 j 而变，然后用 MLE 进行估计（多项）。

但有些解释变量可能既随个体而变，也随方案而变，这种解释变量称为’随方案而变’，既包括同时随方案与个体而变的变量，也包括随方案而变但不随个体而

变的变量（条件），将解释变量不随方案而变的多项L和随方案而变的条件L同时发生就是混合的L。如果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，次假定称为“无关方案的独立性(IIA)’在实践中，如果不同方案之间很类似，则 IIA 假定不一定满足，这是多项，条件，混合共同缺点，此时考虑使用嵌套L 步骤：首先判断数据的基本特征和数据格式，通过描述性统计粗略判断要研究的关系 进行回归 IIA检验 相对风险比率 预测结果

- 实例分析题来自第十二章

由于没有指定参照方案(base outcome)，故命令 mlogit 自动选择观测值最多的方案(即专业人士)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(white)更不可能选择服务业或工匠；但是否白人对于选择蓝领或白领没有显著影响。受教育程度(ed)越高，越不可能选择除专业人士以外的职业。工龄越长(exper)，越不可能选择服务业或蓝领；工龄对于选择工匠或白领无显著影响。

由于 IIA 假定是多项 Logit 模型的前提，下面检验 IIA 假定是否满足。

上表的前四行豪斯曼检验结果显示，去掉四个非参照方案(nonbase alternatives)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”，故上表第五行计算去掉参照方案(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。但由于某种原因，却无法进行 Small - Hsiao 检验。

由以上结果可知，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。这意味着，使用多项 Logit 或多项 Probit 在实际上并无多少区别；只是多项 Probit 的计算时间更长，且无法从几率比角度解释系数估计值，故实践中常使用多项 Logit。

其中，选择项“sepby(id)”表示根据变量 id 的取值来画上表中的横线(默认每隔5个观测值画一条横线)。上表显示，每个旅行团(由变量 id 指定)对应于3行数据，每行对应于一种旅行方式(以变量 mode 表示)。虚拟变量 Train = 1 表示，该行数据对应于乘火车的方案；虚拟变量 Bus = 1 表示，该行数据对应于乘长途大巴的方案；如果 Train = 0 且 Bus = 0，则该行数据对应于自驾车的方案。即使对于同一旅行团，每个方案的总旅行时间、乘车成本也是不同的。另一方面，对于同一旅行团，其家庭收入与旅行团规模则不变。被解释变量 choice 为虚拟变量，表示选择哪一种方案(比如，第1个旅行团选择自驾车，故 car 所对应的那一行 choice = 1，而其他两行 choice = 0)。

尽管数据集中只有152个旅行团，但由于每个旅行团占据3行数据，故实际的样本容量为456(即152 ×3)。下面看一下数据的统计特征。

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。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。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。为此，在上述命令中加上选择项“or”来计算风险比率。

```
. clogit choice train bus time invc,group(id) nolog or
```

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，则选择此方案的概率将乘以0.98，即下降2%。变量 invc 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

上表显示，第 1 个旅行团实际选择自驾车(Car)，而模型预测选择自驾车的概率高达 0.925，而且正好是旅行时间与成本最低的方案。

对于条件 Logit 模型，也可用命令 asclogit 来估计。

上表显示，家庭收入(hinc)越高，越不倾向于选择火车；但对于选择长途大巴则无显著影响。旅行团规模(psize)没有显著影响。旅行时间(time)与乘车成本(invc)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。上表并不汇报准 R^2，但可以很容易地手工计算。从上表可知，该模型的对数似然函数为 -77.504 846，下面估计一个只含常数项的模型。

上表显示，只含常数项模型的对数似然函数为 -160.001 72，故可计算准 R^2 如下。

```
. dis (160.00172 -77.504846)/160.00172
. 51559992
```

从上表可知，对于方案 Train 与 Bus，变量 Type 的取值为 public；而对于方案 Car，变量 Type 的取值为 private。在数据集中，有两个不随方案而变的解释变量，即 hinc 与 psize。为了演示的目的，假设 psize 的系数只随树干方案而变(仅取决于是公共交通还是私人交通)[①]，而 hinc 的系数可随树枝方案而变(对于 Train，Bus，Car 各不相同)。

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(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)。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。

需要注意的是，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；而有些多值选择模型不存在清晰的嵌套结构。如果 IIA 假定不成立，但又没有清晰的嵌套结构，可考虑