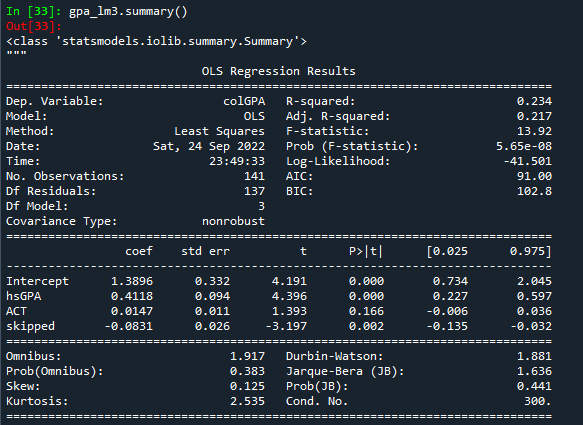
## 2 回顾中用OLS实现模型的求解以及CLM假设之下OLS解的一些性质

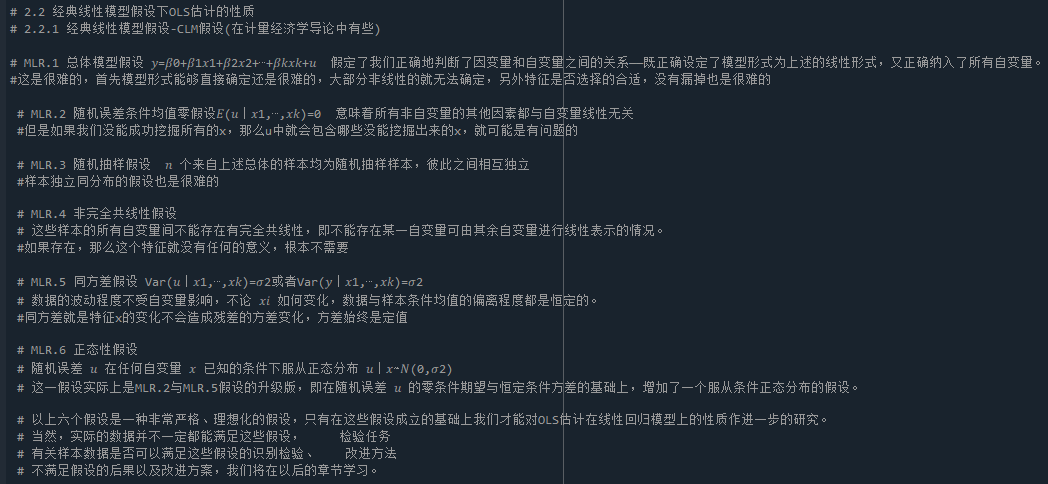
### 1.求解：

gpa\_lm3=sm.formula.ols('colGPA~hsGPA+ACT+skipped',data=gpa1).fit()

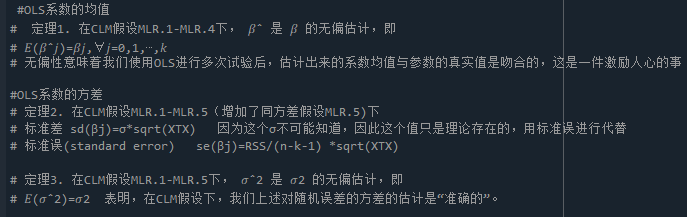
gpa\_lm3.summary()

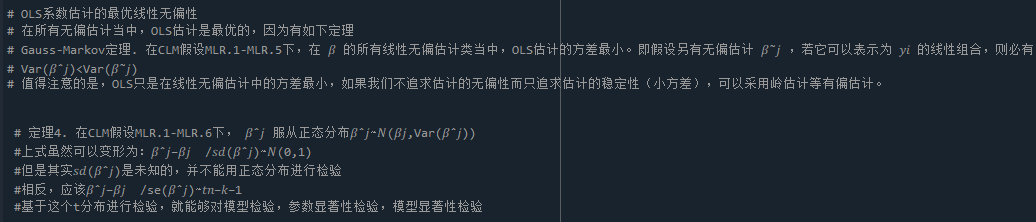


### 2. CLM的6个假设



### 3.基于CLM的6个假设，OLS估计值具有的几个性质





## 3 假设检验，参数检验问题

### 3.1 t检验

#### 3.1.1 用途：

单参数检验，也就检验系数的显著性，原假设为0，备择假设不为0

参数的线性组合假设，也就是检验系数是否相等，原假设相等，备择假设不相等

#### 3.1.2 实际操作

双边检验：

计算，并与一定置信水平α下的t分布的接受域（接受域就是α/2到1-α/2之间）比较

单边检验：

备择假设是小于的时候，就关注T值和α分位点，

备择假设是大于的时候，就关注T值和1-α分位点

#### 3.1.3 p值

为什么要用p值？

因为上面本质上，实际上为了避免C的计算，用的与一定置信水平α下的t分布的接受域的边界点比较，都是明显和置信水平α相关的，用p值就能够直接看出什么置信水平下接受或者拒绝

双边检验：

summary中的p值，正是系数0值双边检验的p值，可以直接使用

单边检验：

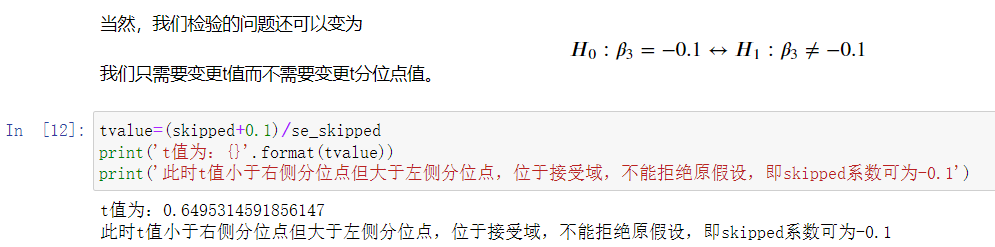
由于双边检验p值是对单边检验p值乘两倍得来的，我们要根据t值是否大于0来选择左/右尾累积概率，若小于0，则选择左尾；反之右尾。

sf:右尾累积概率

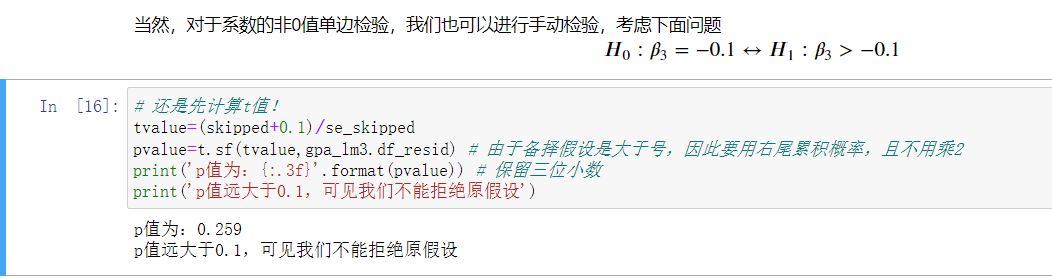
cdf:左尾累积概率

#### 3.1.4 系数的非0检验

如果用t检验：

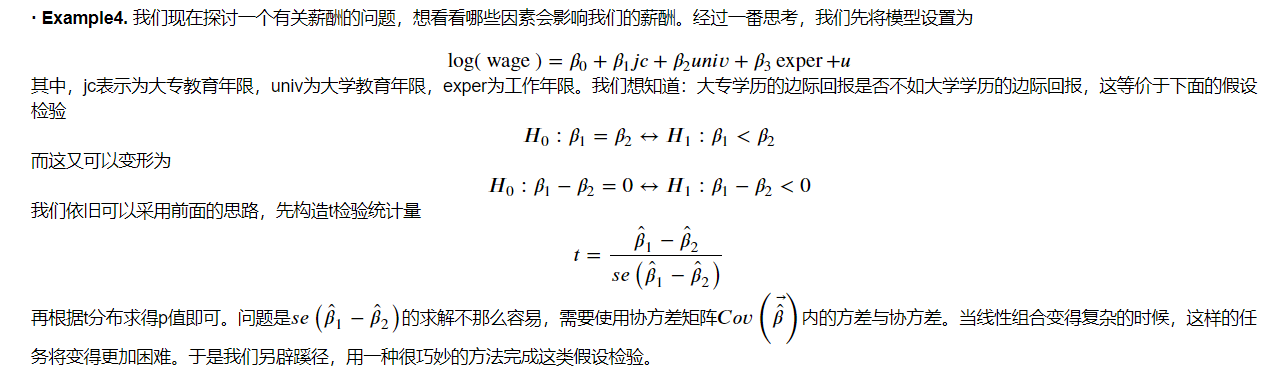


如果用p值：

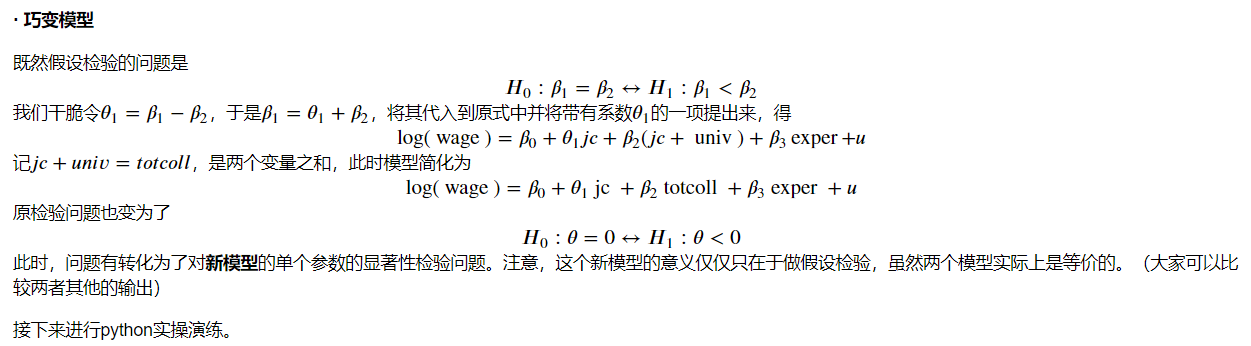


#### 3.1.5 不是单个参数，而是参数线性组合的检验

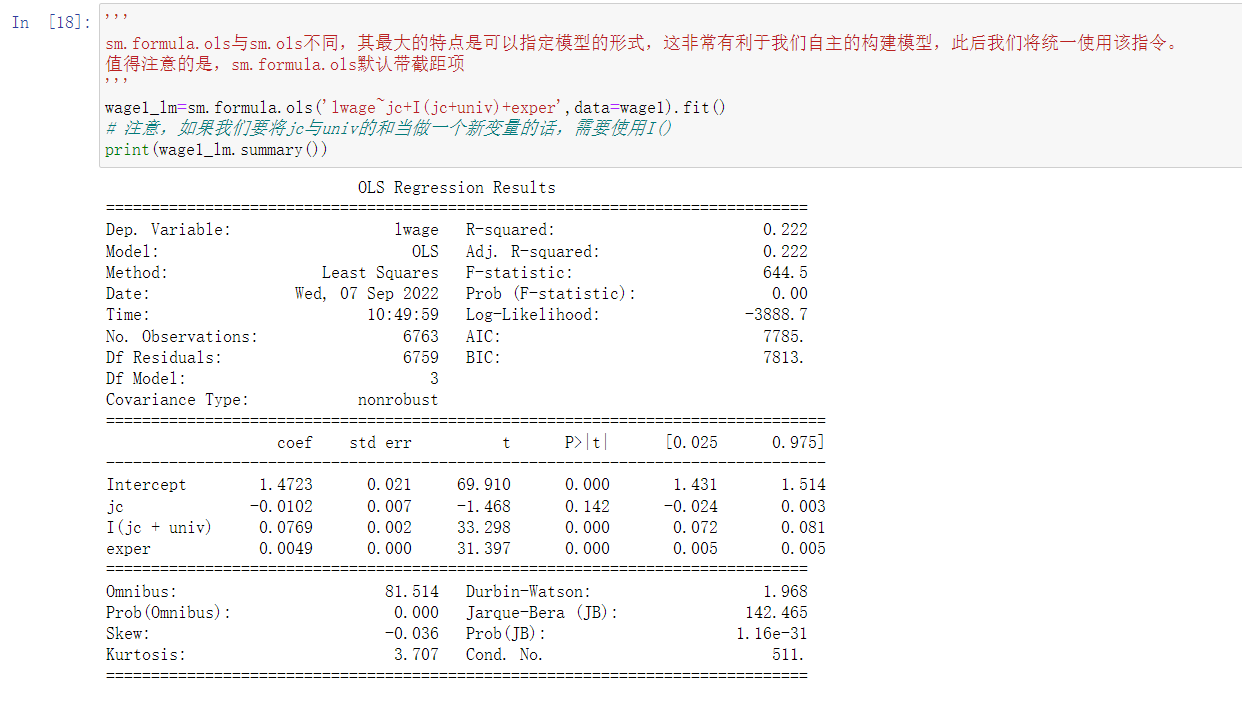
方式一：从数学理论出发：



方式二：通过变形使问题转换成上面的单参数问题



对于方式二，也很容易求解：



#### 3.1.6 双侧p值进行单侧0值检验

双侧p值进行单侧0值检验的小技巧。在summary中，jc的t值小于0，说明它后面的双侧p值是使用左侧累积概率乘两倍得来的，而在本例中我们的备择假设是小于0，p值也应当是左侧累积概率，因此我们只需将报告表中的p值除以2即可。而如果t值小于0，但是备择假设却大于0，那么无需思考，p值一定大于0.5，我们肯定不能拒绝原假设。

双侧检验的P值为检验统计量X 落在样本统计值C 为端点的尾部区域内的概率的2 倍：P = 2P{ X > C} (当C位于分布曲线的右端时) 或P = 2P{ X< C} (当C 位于分布曲线的左端时)

### 3.2 F检验

注意，三个参数做联合显著性检验**完全不等价于**三个参数分开做显著性t检验！如果我们是出于联合检验的目的但是却做了分开检验，将大大增加拒真概率。由于无法分开始用t检验进行联合检验，我们需要一种新的检验方法——F检验。

#### 3.2.1 类别：

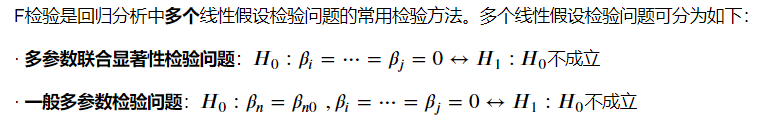
F检验是回归分析中**多个**线性假设检验问题的常用检验方法。多个线性假设检验问题可分为如下：

· **多参数联合显著性检验问题**：𝐻0:𝛽𝑖=⋯=𝛽𝑗=0↔𝐻1:H0:βi=⋯=βj=0↔

H1: 𝐻0H0不成立

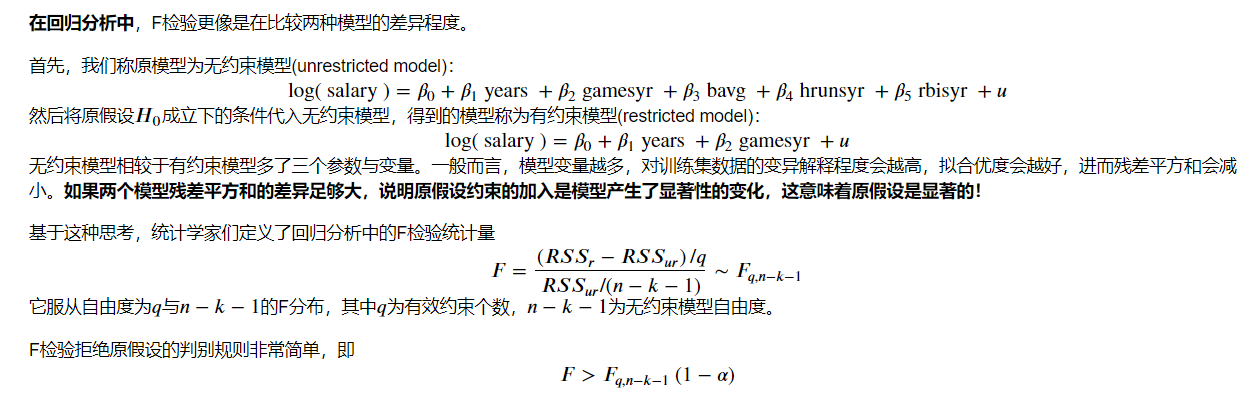
· **一般多参数检验问题**：𝐻0:𝛽𝑛=𝛽𝑛0,𝛽𝑖=⋯=𝛽𝑗=0↔𝐻1:H0:βn=βn0,βi=⋯=βj=0↔

H1: 𝐻0H0不成立



#### 3.2.2 ****多参数联合显著性检验问题（也就是检验多个参数是否同时为0）****

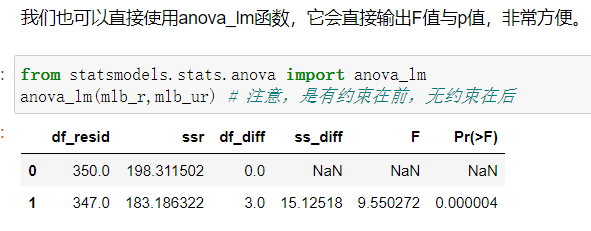
基本理论



手动计算



软件计算



#### 3.2.3 ****一般多参数检验问题（也就是检验某些参数是否不为0）****



由于我们做的是0值单边假设，因此可以通过summary汇总表中的p值判断。female的t值小于0，因此其p值采用的是左尾累积概率；而我们的假设是小于0假设，也采用的是左尾累积概率，因此我们只需要将报告表中的p值除以2即可。显然，female的p值在保留三位小数的前提下依旧为0.000，因此它除以2后一定也为0.000，我们可以拒绝原假设，职场上男女薪资存在不平等现象。

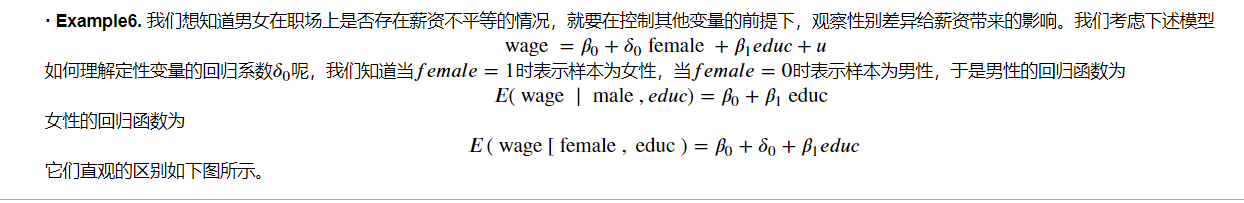
## 4. 广义线性回归，包括类别变量，以及交互性质等等

自变量不仅可以是一次的连续变量，还可以是一种**定性变量**，也可以是某个**变量的函数**，如二次项𝑋2X2、对数项𝑙𝑜𝑔(𝑋)log(X)。这是因为，所谓的线性回归模型，线性关系并不是指代被解释变量𝑦y与解释变量𝑋X之间的关系，而是指回归函数相对于**回归系数**是线性的。

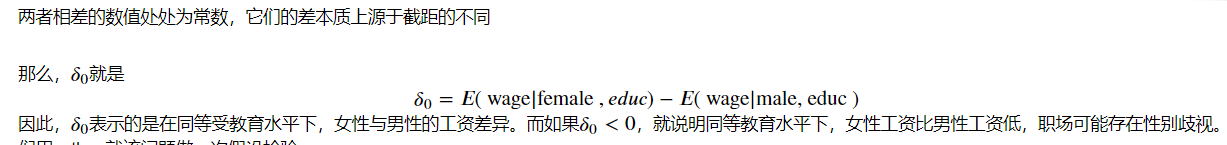
### 4.1 类别变量（classes=2）处理

#### 4.1.1 没有交互效应时候类别变量

处理方式其实就是引入0-1虚拟变量

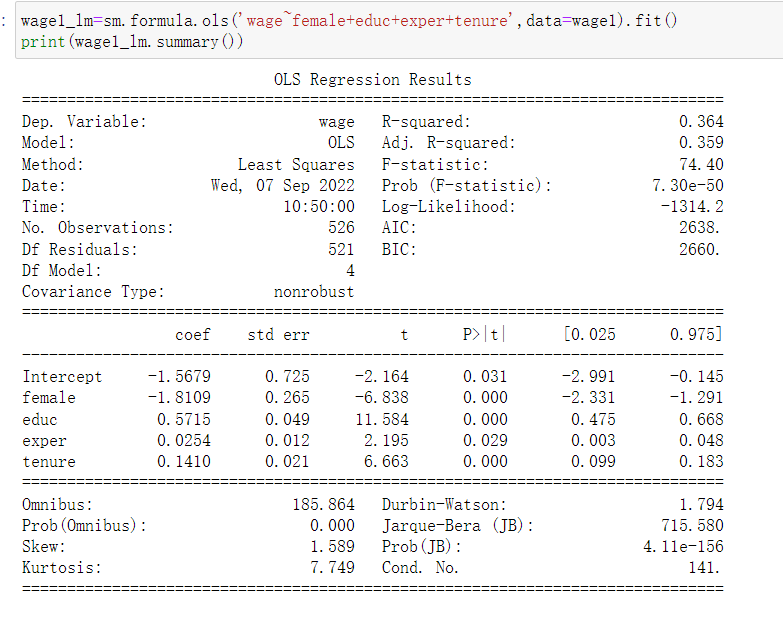


可见这个模型的意思就是教育的影响都是β1，但是男性和女性的截距是不同的



要检验的其实就是  
𝛿0<0成立与否

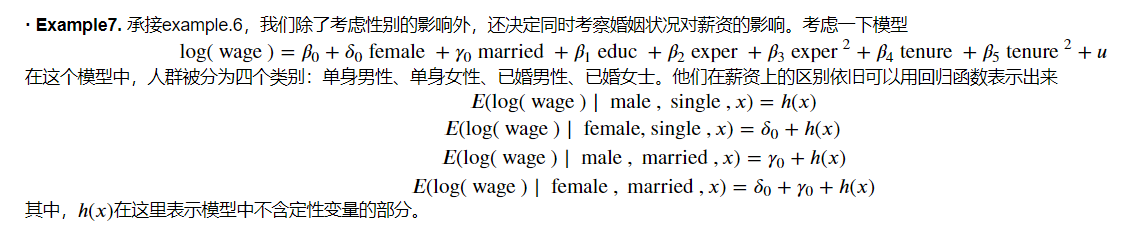
结果是：



可见female的t值是小于0的，然后双侧p值0，除以2之后还是0，因此说明拒绝原来假设

#### 4.1.2 有交互效应时候的类别变量

不考虑交互项时候存在的问题

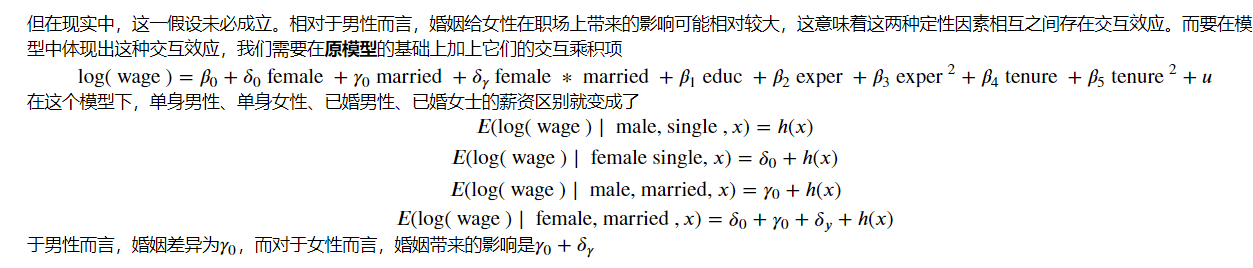


但是其实下面的模型中不存在交互项：

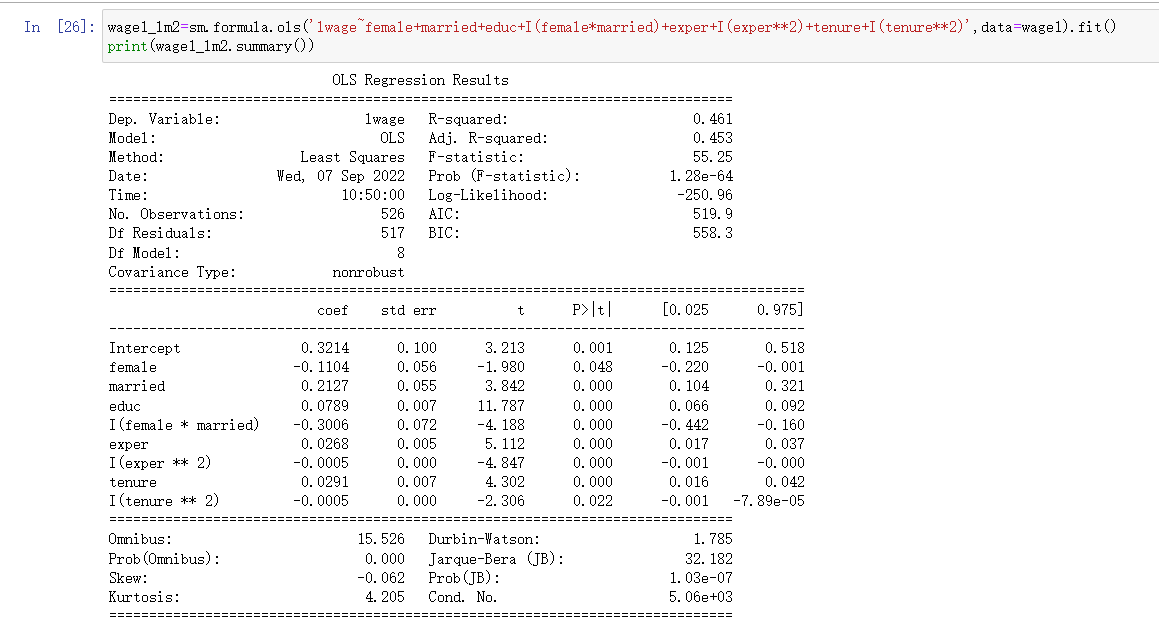


则不论是未婚还是已婚，性别差异都是𝛿0；不论是男性还是女性，结婚与否的差异都是𝛾0，肯定是不对的

考虑交互项时候的模型：



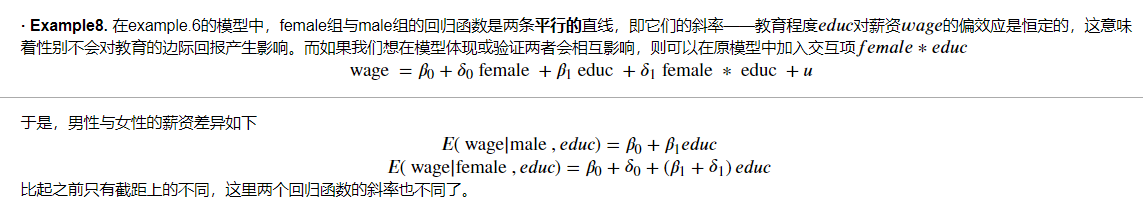
因此看𝛿𝛾是否等于0，如果等于0 ，这说明其实不存在交互影响，婚姻差异对男女性来说是相同的

  
明显𝛿𝛾的p值接近于0，则拒绝原来假设，性别与婚姻状况确实存在交互效应；

于男性而言，已婚人士平均工资比未婚人士的差异其实就是𝛾0=0.2127

对于女性而言，已婚人士比未婚人士工资的差异是𝛾0+𝛿𝛾=0.2127-0.3006也就是-8.8%，可见结婚对男性和女性的影响非常不一致！

#### 4.1.3 类别变量与数值变量的交互效应



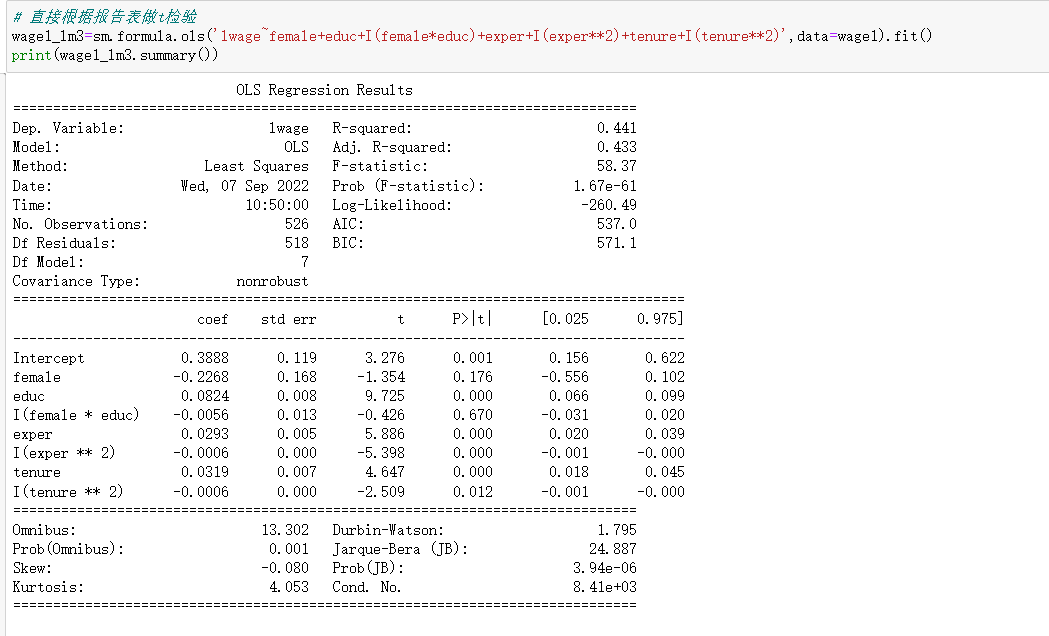
检验1：检验男性与女性的边际教育回报是否相同。

这等价于检验假设

𝐻0:𝛿1=0↔𝐻1:𝛿1≠0

也就是说educ前面的系数是否相等

这个检验使用t检验即可：



可见p=0.67，是很大的，也就是说两性之间教育边际回报相等的假设是不能被拒绝，即两性边际回报可以认为是相等的

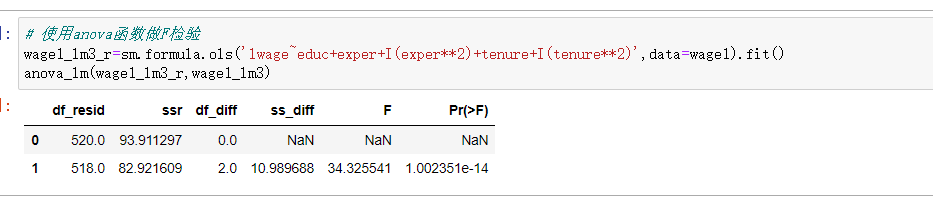
检验2：检验男性与女性的平均工资是否存在性别差异。

这等价于检验假设

𝐻0:𝛿0=𝛿1=0↔𝐻1:∃𝛿𝑗≠0,𝑗=0,1

也就是说两个期望值是否完全相等，如果完全相等的时候𝛿0=𝛿1=0必须成立

这是两个参数，使用F检验



可见这个p几乎等于0，因此拒绝原假设，两性存在薪资差异

上述两个检验在一起说明，引起两性收入不平等的原因并非来自教育程度。

### 4.2 多类别变量（classes>2）处理

#### 4.2.1 处理方式与注意问题

不能使用整数编码，因为不同分类之间的差异完全取决于取值之间的差

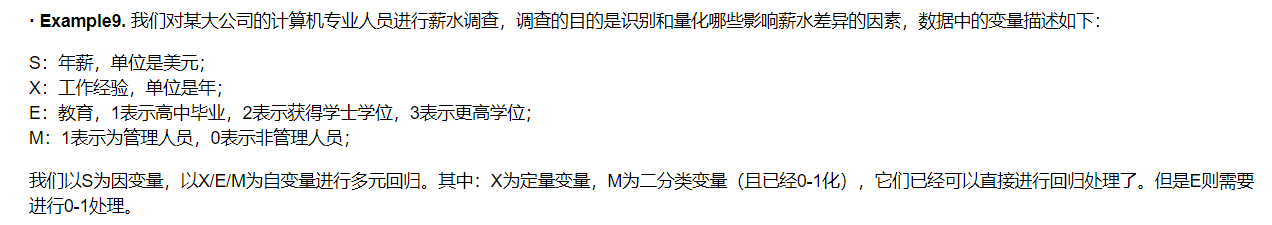
而是用多个二值虚拟变量来表示多分类定性变量。**具体的，如果一个变量有n个类别，则需要定义n-1个虚拟变量表示它**。

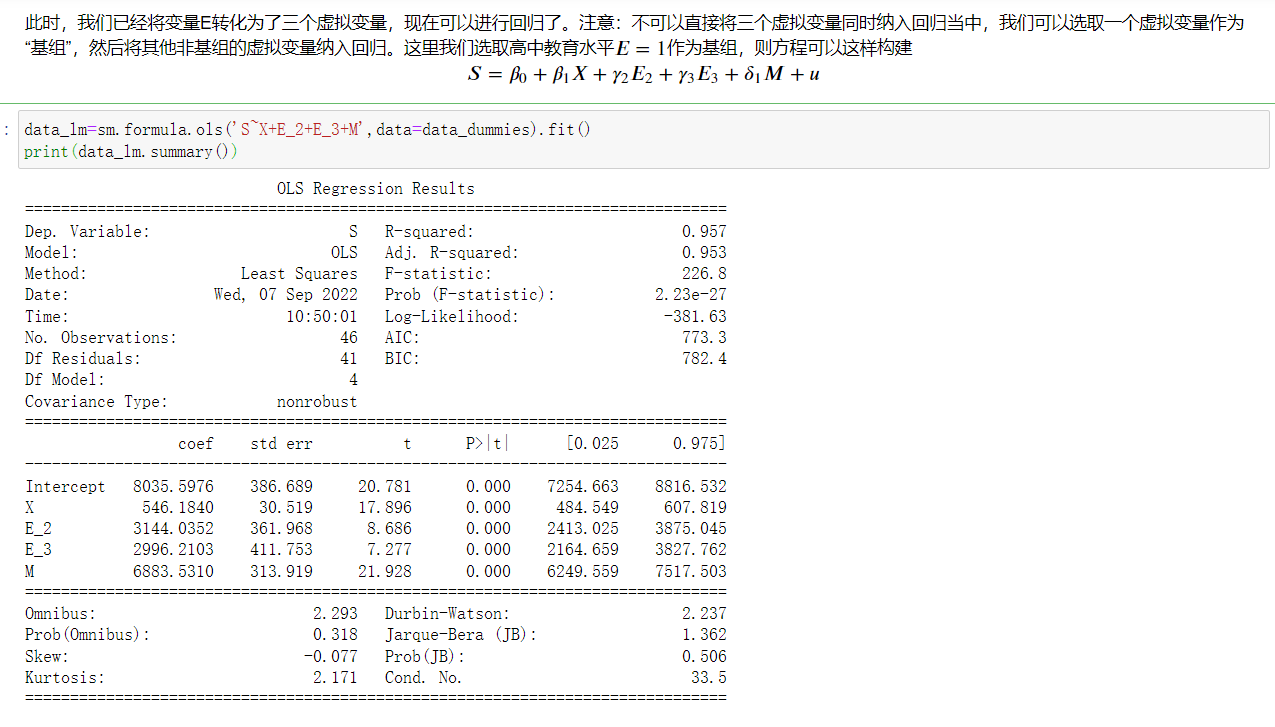
为什么要使用n-1个呢？因为n个自变量存在完全共线性，违背了CLM假设中的MLR.4，使得模型完全失效。

#### 4.2.2 如何将整数编码转换成独热编码？

get\_dummies这一函数会自动变换所有具有对象类型（如字符串）的列，但是如果某列的变量是数值型变量（哪怕它实际上是分类变量），它将不会为该列创建虚拟变量，除非我们将该列的数据类型从数值转化为字符串。

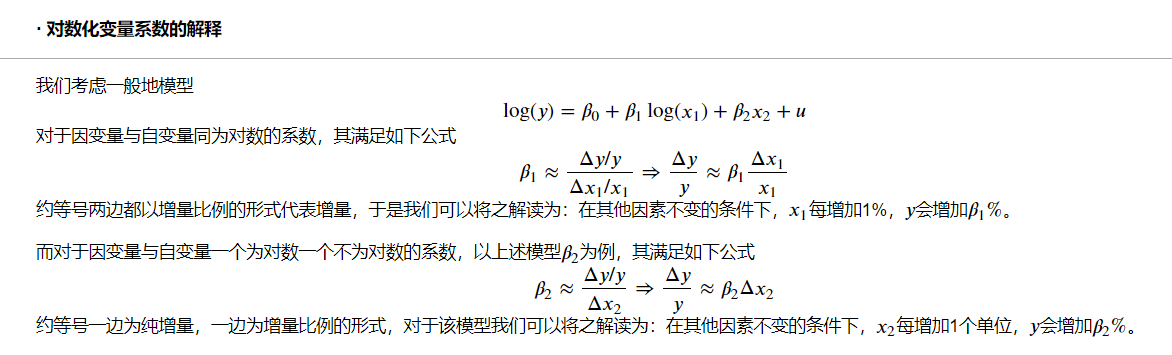
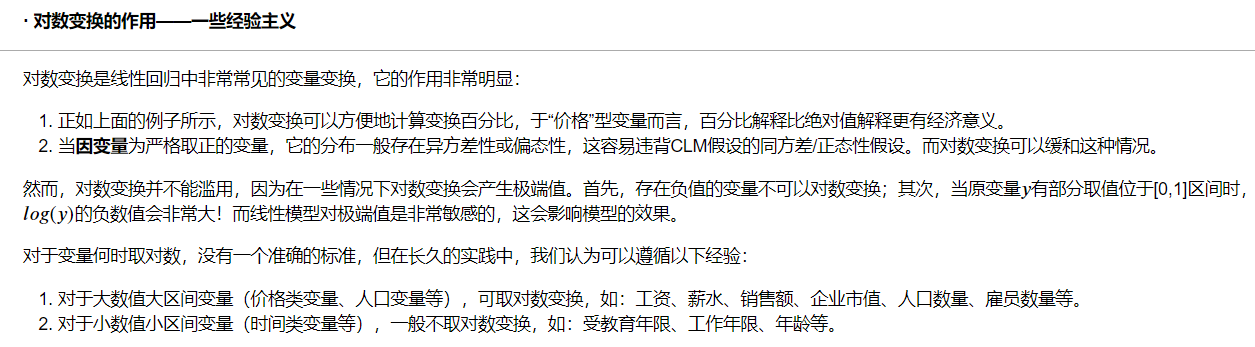
#### 4.2.3 实例





### 4.3 变量以函数形式呈现（对数化，二次项化，交互项）

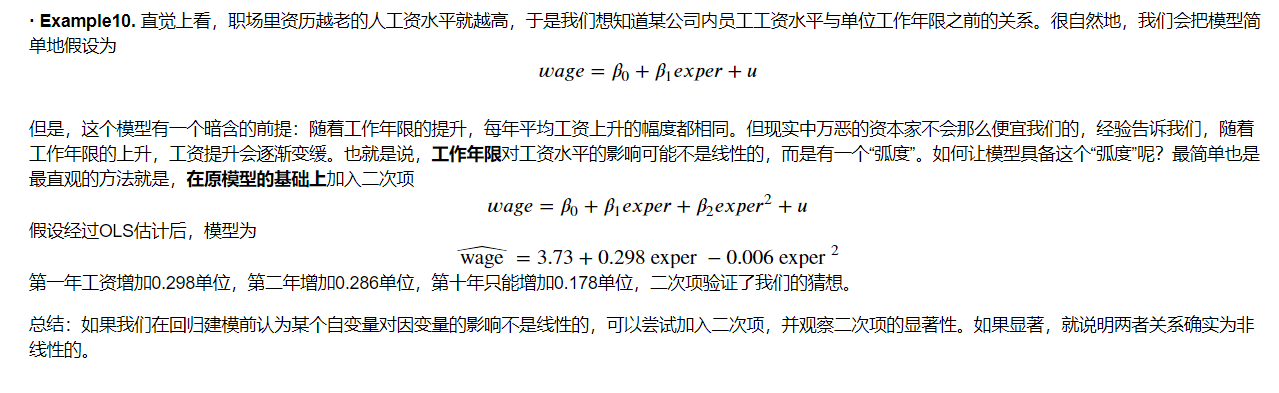
#### 4.3.1 对数化

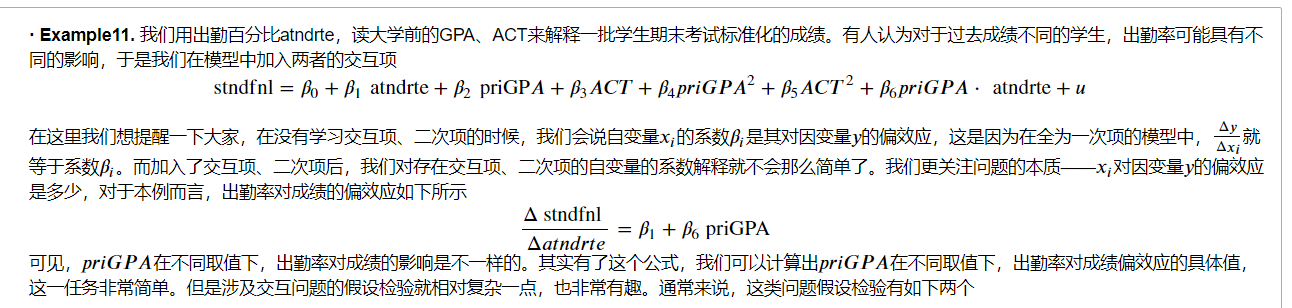
也就是说正数，但是存在偏态分布，可以使用对数化使之变成正态分布

#### 4.3.2 二次项

就是那些非线性变化的，可以在一定程度内使用二次项进行拟合，当然这个数据不能像exp那样剧烈波动



#### 4.3.3 交互项

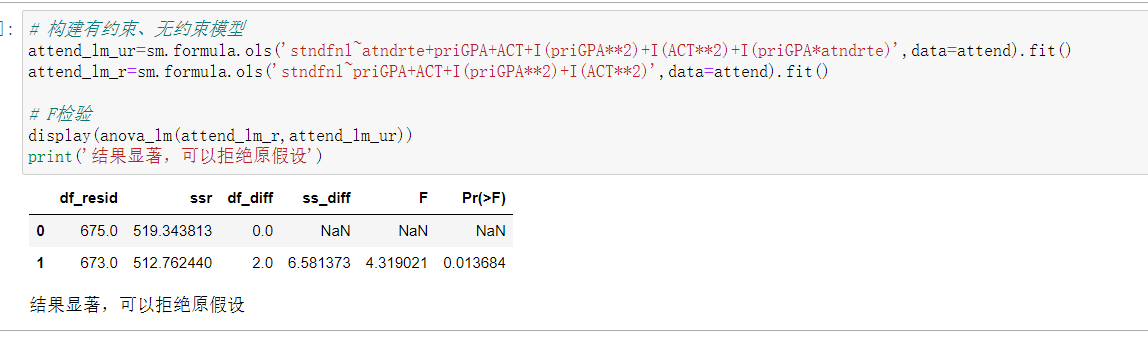


引入了交互项之后，可见标准成绩对出席率的偏导数不是常数，而是跟之前的GPA成绩相关了

检验1：出勤率对成绩的影响是否显著？

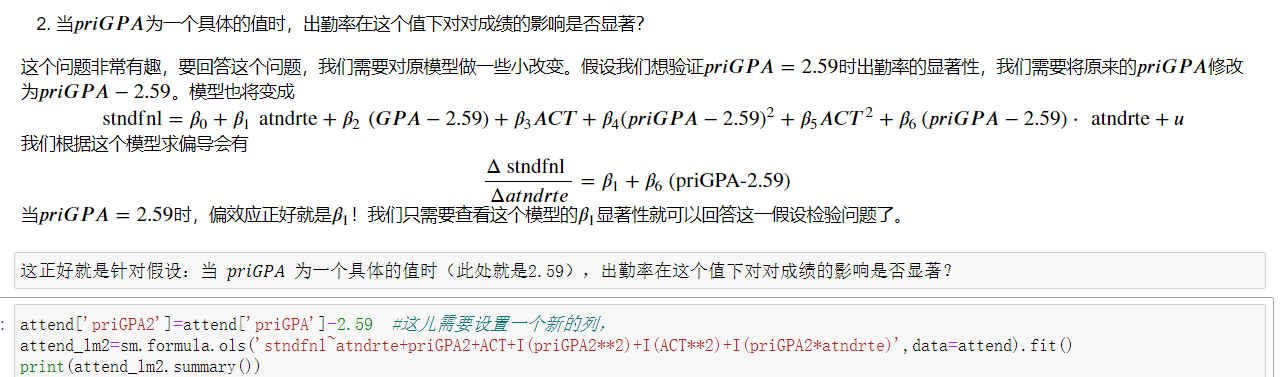
𝐻0:𝛽1=𝛽6=0↔𝐻1:𝐻0不成立

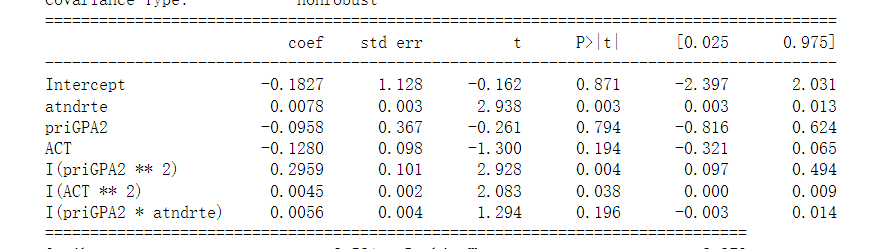
原假设就是没有影响的时候，那么偏导数就是0，自然两个参数都得是0



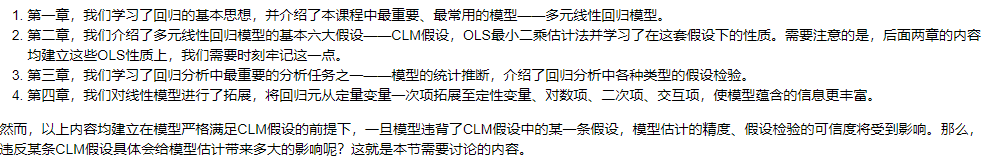
这个p值很小，可见拒绝原来的假设

检验2：当𝑝𝑟𝑖𝐺𝑃𝐴priGPA为一个具体的值时，出勤率在这个值下对对成绩的影响是否显著？（这个还是第一次见）



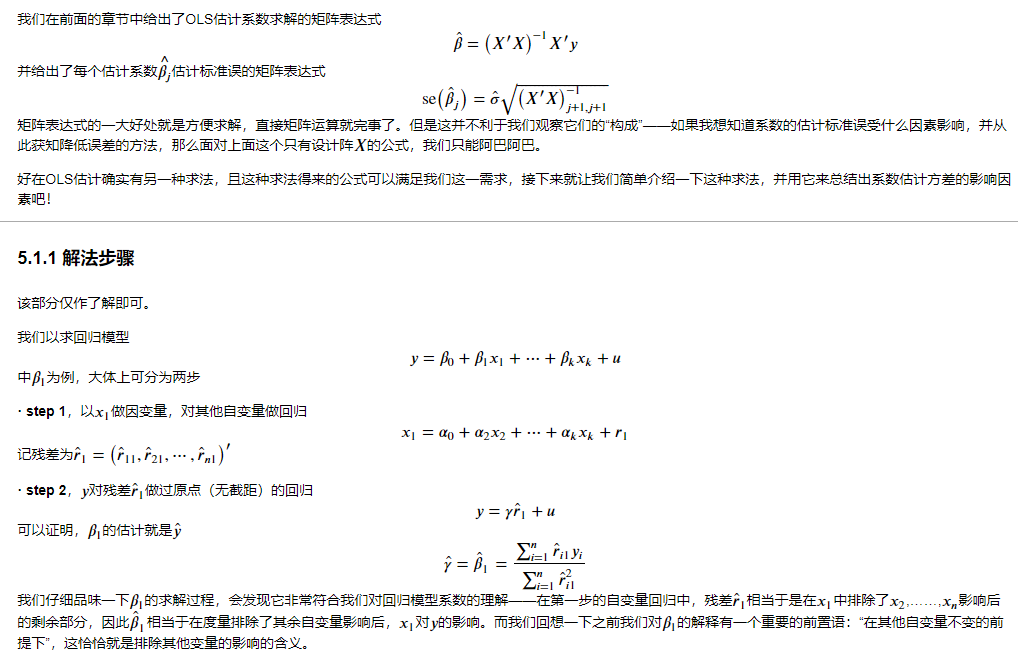
结果显示β1的p值很小，因此拒绝原假设

## 5. CLM假设误差分析

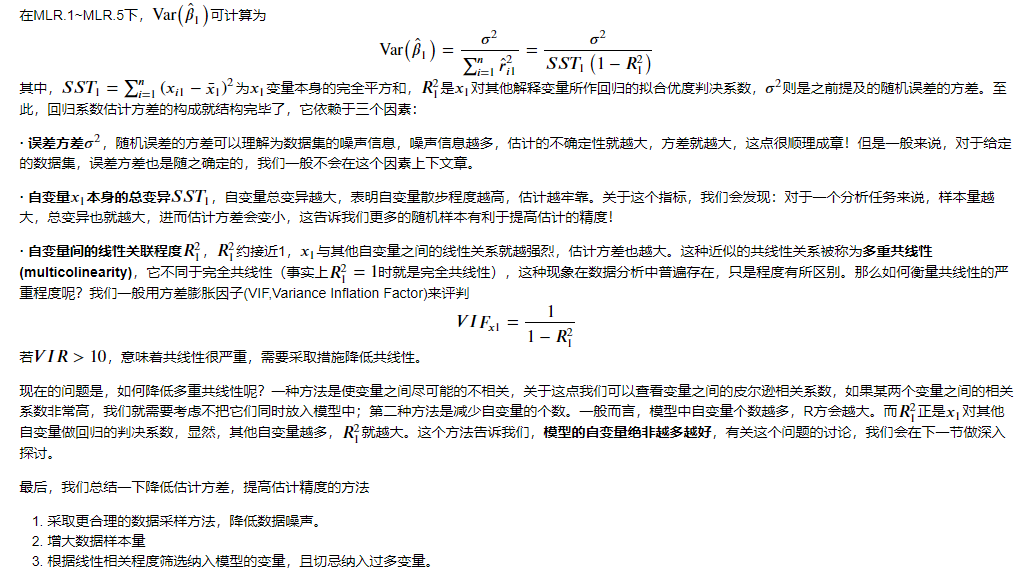


### 5.1 参数的另外一种求解方法

#### 5.1.1 解法原理



#### 5.1.2 估计系数的方差构成，以及解决办法



**误差方差**𝜎2σ2，随机误差的方差可以理解为数据集的噪声信息，噪声信息越多，估计的不确定性就越大，方差就越大，这点很顺理成章！但是一般来说，对于给定的数据集，误差方差也是随之确定的，我们一般不会在这个因素上下文章。

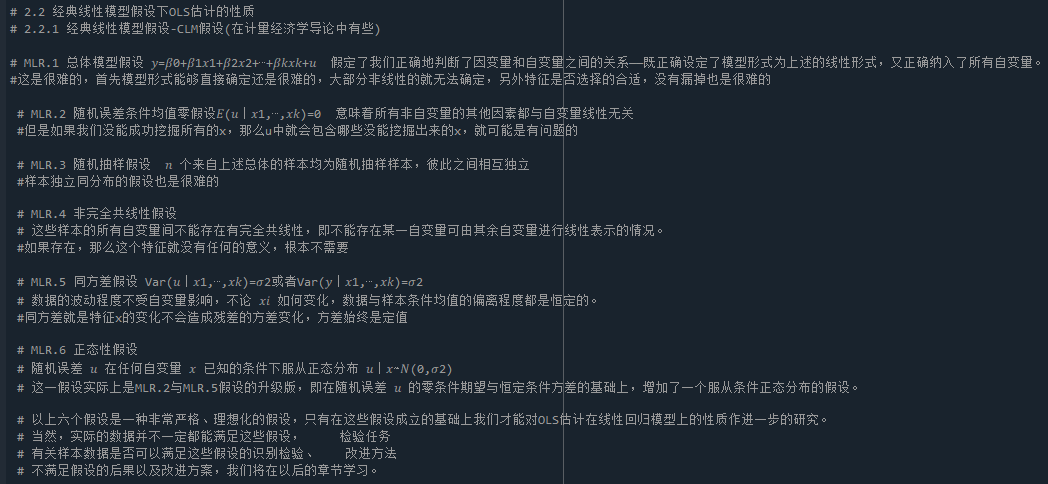
**· 自变量**𝑥1x1**本身的总变异**𝑆𝑆𝑇1SST1，自变量总变异越大，表明自变量散步程度越高，估计越牢靠。关于这个指标，我们会发现：对于一个分析任务来说，样本量越大，总变异也就越大，进而估计方差会变小，这告诉我们更多的随机样本有利于提高估计的精度！

**· 自变量间的线性关联程度**𝑅，𝑅越接近1，𝑥1与其他自变量之间的线性关系就越强烈，估计方差也越大。这种近似的共线性关系被称为**多重共线性(multicolinearity)**，它不同于完全共线性（事实上𝑅21=1时就是完全共线性），这种现象在数据分析中普遍存在，只是程度有所区别。那么如何衡量共线性的严重程度呢？我们一般用方差膨胀因子(VIF,Variance Inflation Factor)来评判

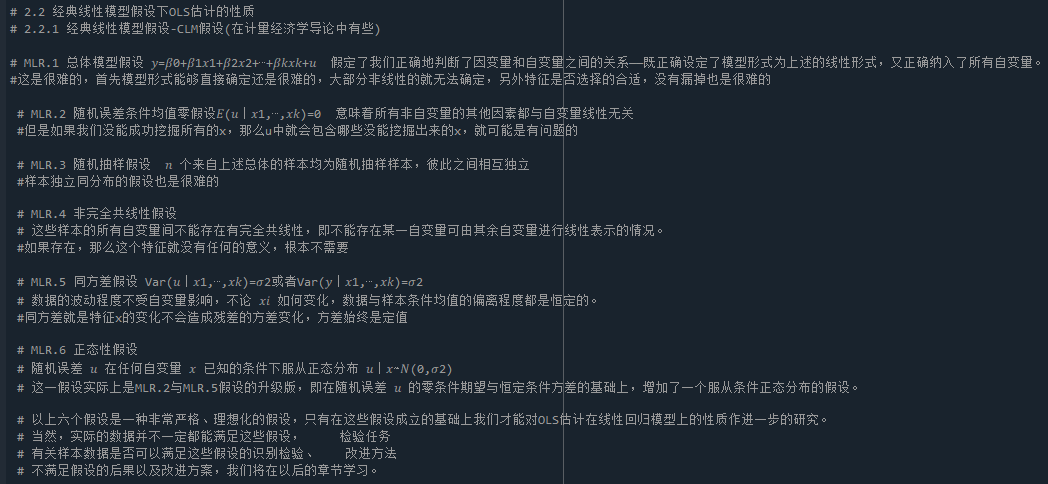
若𝑉𝐼𝑅>10，意味着共线性很严重，需要采取措施降低共线性。

现在的问题是，如何降低多重共线性呢？一种方法是使变量之间尽可能的不相关，关于这点我们可以查看变量之间的皮尔逊相关系数，如果某两个变量之间的相关系数非常高，我们就需要考虑不把它们同时放入模型中；第二种方法是减少自变量的个数。一般而言，模型中自变量个数越多，R方会越大。而𝑅21正是𝑥1x1对其他自变量做回归的判决系数，显然，其他自变量越多，𝑅21就越大。这个方法告诉我们，**模型的自变量绝非越多越好**，有关这个问题的讨论，我们会在下一节做深入探讨。

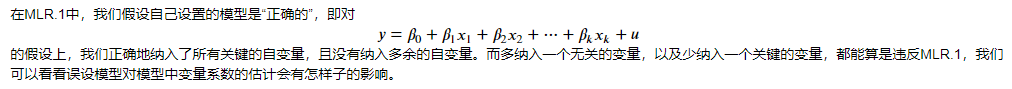
### 5.2.0 首先回顾6个假设



### 5.2 模型误设的误差分析——违反MLR.1的后果



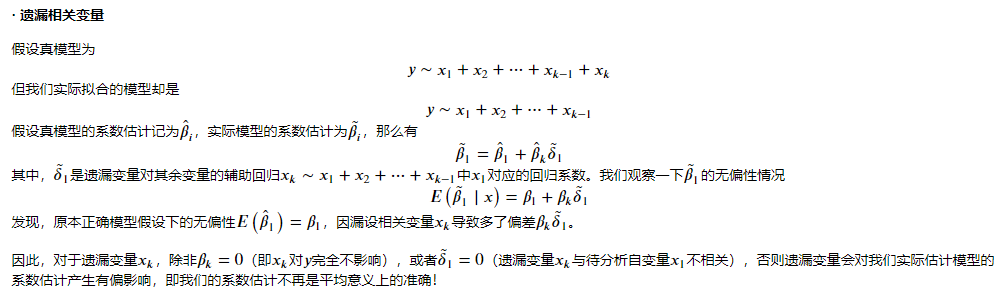
#### 5.2.1 如何理解模型误设



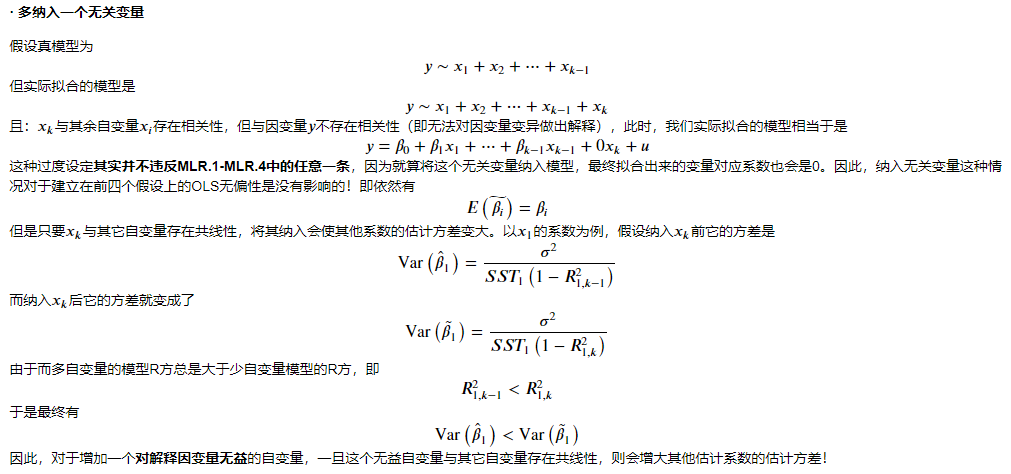
因为我们现在只是分析线性模型，因此无法改变模型的形式，只是说是不是变量引入出现了问题

#### 5.2.2 模型误设的后果（要么偏差大方差小，要么偏差小方差大）

**遗漏相关变量会造成有偏估计**



**多纳入一个无关变量虽然仍然是无偏估计，但是方差会变大**

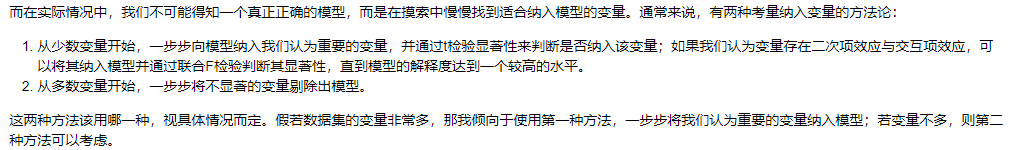


#### 5.2.3 变量选择结论



但是到底选择哪些合适的变量呢？

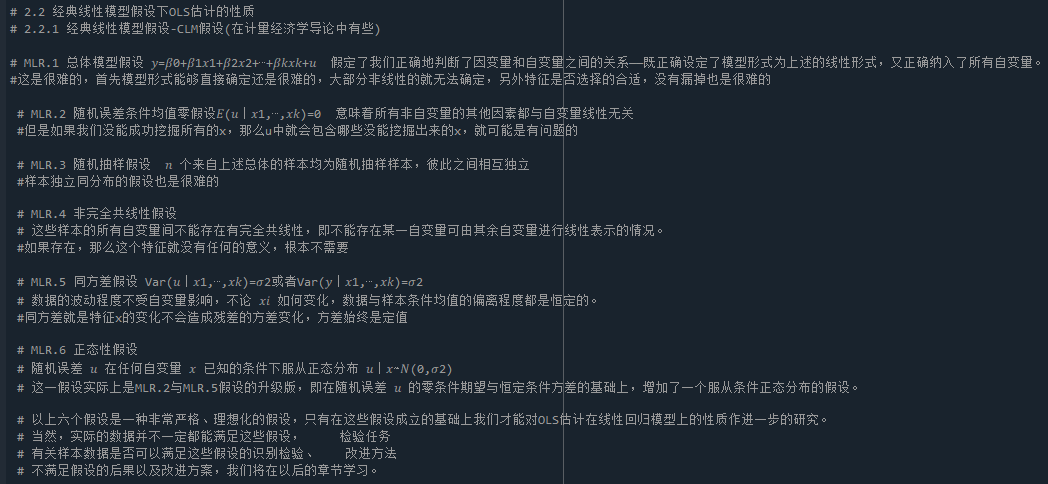
存在向前和向后两种方法



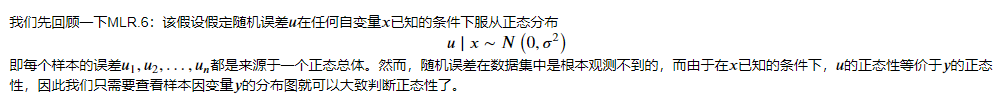


例如说使用RF，就是不断地选择特征，其实最后可以得到特征重要性这个度量

### 5.3 模型不满足正态性的分析——违反MLR.6的后果

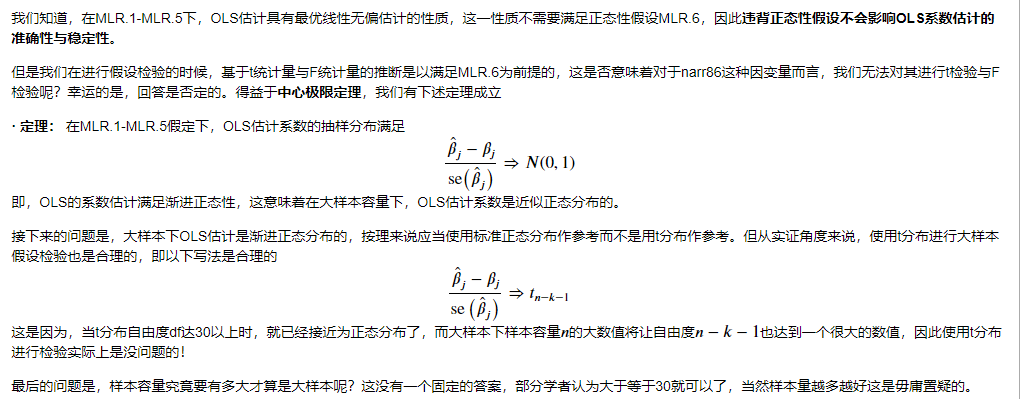


#### 5.3.1 如何理解与观测正态性假设



然而，因变量y不为正态分布的情况是非常常见的，

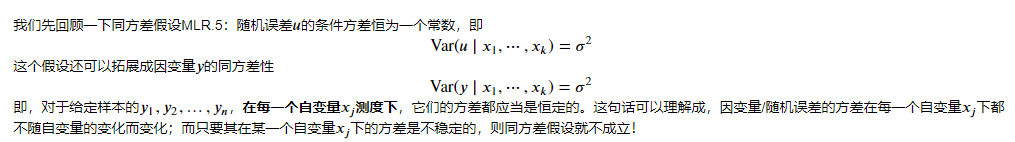
#### 5.3.2 违背正态性假设的后果（只要数据量足够多就不是问题）



上面说明，只要数量够多，本来理论上就可以得到最优线性无偏估计，而且使用t检验、F检验也是完全没问题的

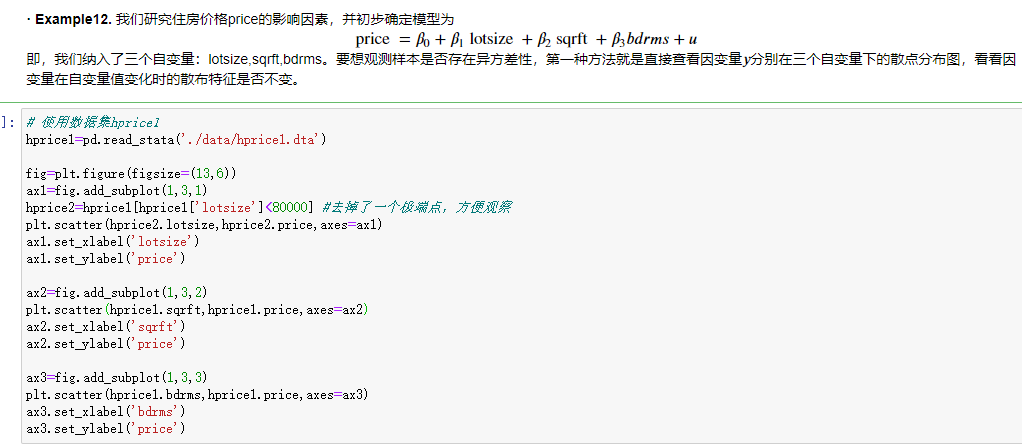
### 5.4 异方差的分析——违反MLR.5的后果

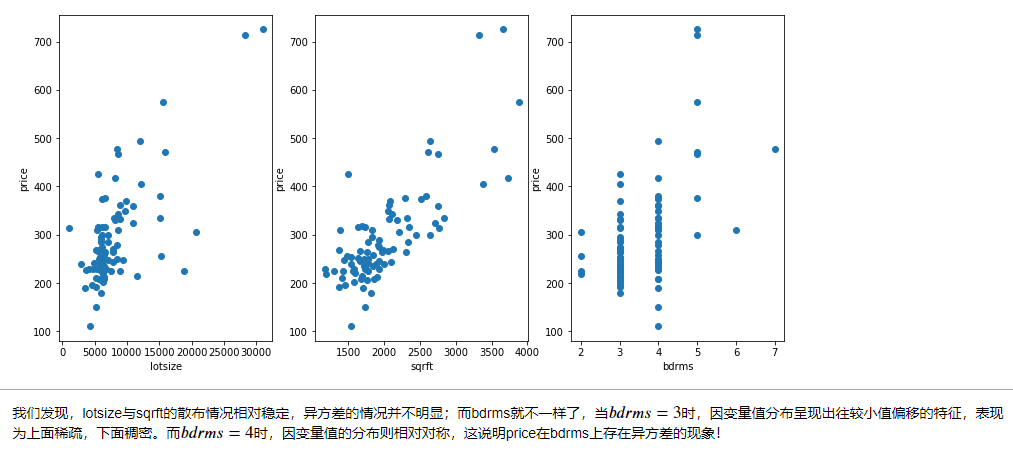
#### 5.4.1 如何理解与观测异方差性



一个例子说明：

方法1：



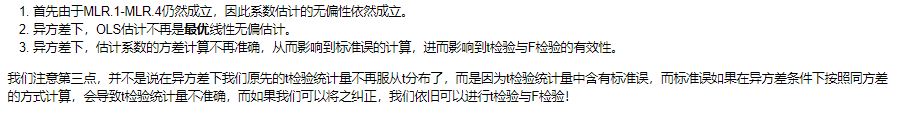


左边两个变量不好看出来，虽然变量变化的时候，标签的均值也在变动，但是不知道标签的方差是否发生了变化，不过dbrms能够很明显说明，3的时候呈现明显的异方差，而dbrms=4的时候却是同方差

方法2：

另一种观测异方差的方法是先进行OLS拟合，并以0为参考基准，观测残差在三个自变量上的表现。

#### 5.4.2 违背同方差假设的后果（方差计算不准确，则t检验和F检验会出现问题）



异方差导致标准误等等无法给出（因为此时标准误就不是一个定值了），那么基于标准误的检验就会出现问题

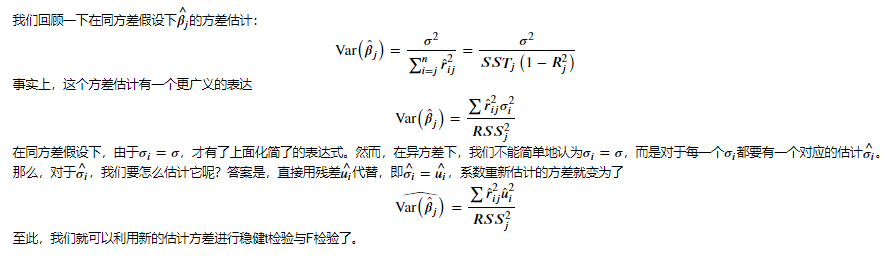
## 6. 异方差下的回归分析（就是用来解决上面的5.4节的问题）

异方差导致的两个后果：OLS不再最优、t检验与F检验不再稳健。

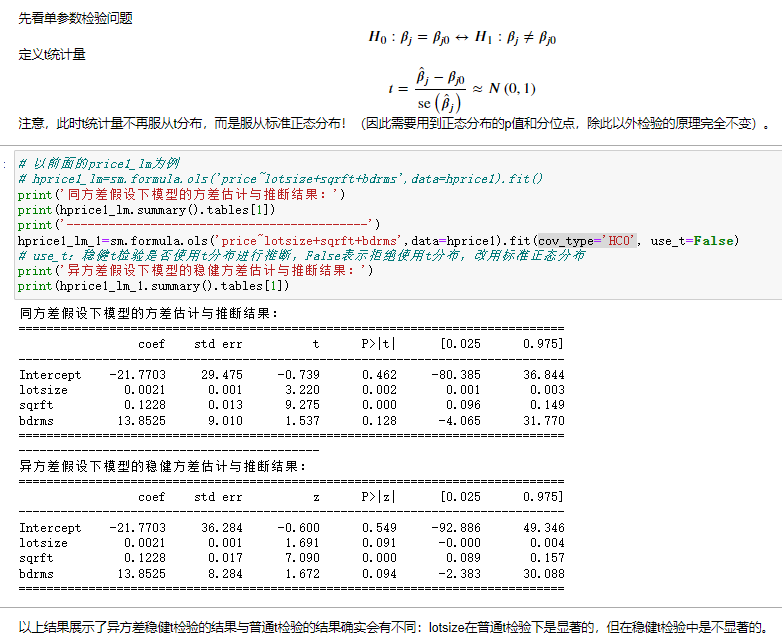
### 6.1 异方差稳健的t检验与F检验

#### 6.1.1 重新估计方差

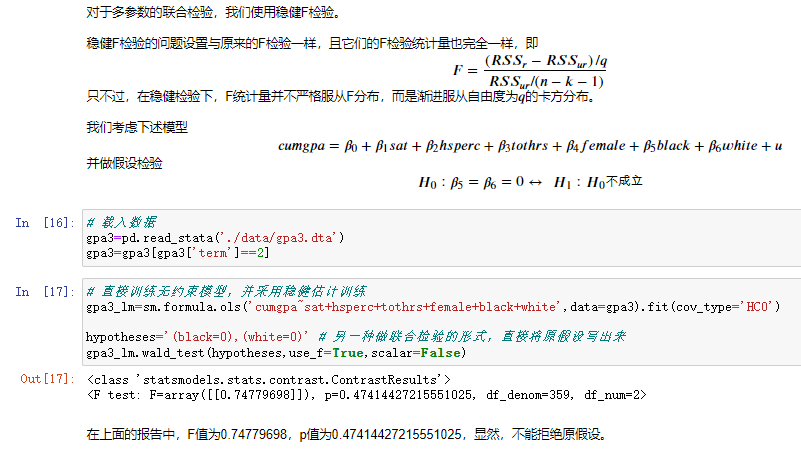
方差不再是同方差假设下的常数，怎么办？



#### 6.1.2 稳健t检验（t统计量不再服从t分布，而是服从标准正态分布！）



#### 6.1.3 稳健F检验（在稳健检验下，F统计量并不严格服从F分布，而是渐进服从自由度为𝑞q的卡方分布）



### 6.2 异方差的诊断

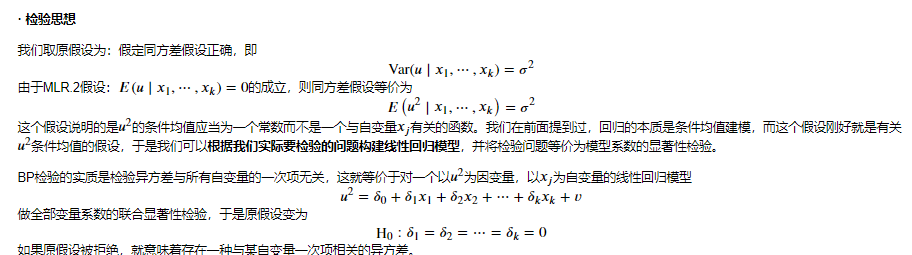
通过散点图的方法粗略地判断模型是否存在异方差，但是这样的判断方法过于主观

方差是某些自变量的函数𝜎=𝜎(𝑥)，就可以采用广义OLS估计法（也就是GLS）进行更有效地估计。

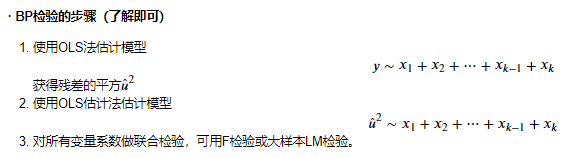
关于检验方差是否与自变量无关的方法，最常用的便是BP异方差检验与White异方差检验。

#### 6.2.1 BP异方差检验（是否存在与变量的一次项相关的异方差）

思想:



流程：

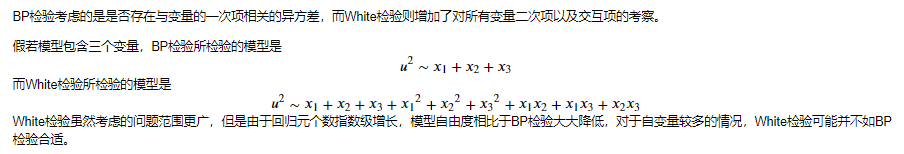


实践：

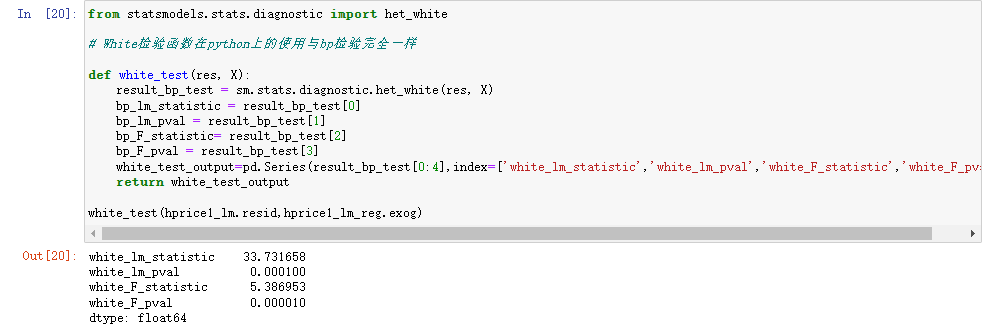


#### 6.2.2 White异方差检验（是否存在与变量的一次项、二次项相关的异方差）

基本原理：



实践：



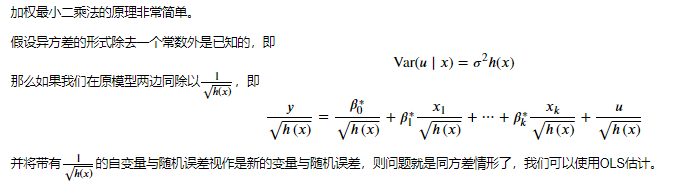
### 6.3 广义最小二乘法（加权最小二乘估计WLS和可行的广义最小二乘估计FGLS）

根据异方差的具体形式使用广义的最小二乘法，对模型进行重新估计。“根据异方差的具体形式”，是指异方差可以用自变量的函数被表达出来。

其一，如果异方差可以用自变量的函数被表达出来，我们就可以是加权最小二乘估计WLS；其二，如果由于函数形式复杂而无法被判断出来，我们则使用可行的广义最小二乘估计FGLS。

#### 6.3.1 加权最小二乘法（要主观判断hx的形式，这个方法没什么意义）

基本思想（但是这个hx好像还是很难确定下来的吧）：整个思想就是hx知道，就除以它，剩下的红框中的方差当然就是同方差的了。

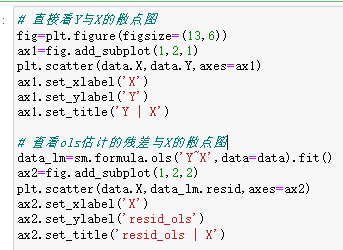


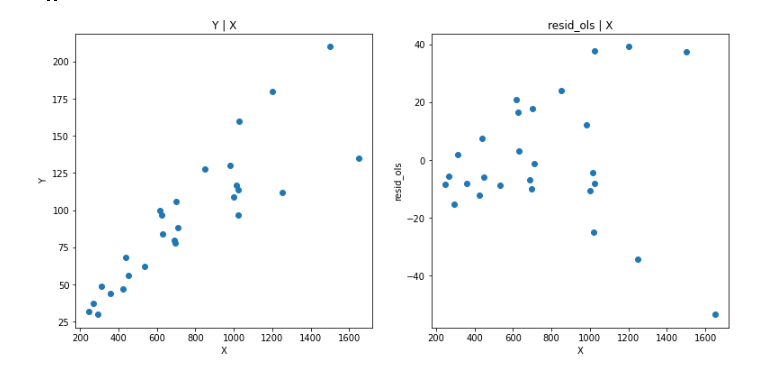
实践：

实践1：检验异方差

采用定性观察散点图和定量BP、white检验的方式，来检验异方差

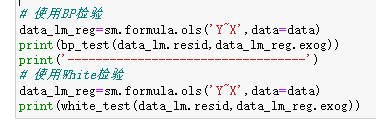
（1）散点图

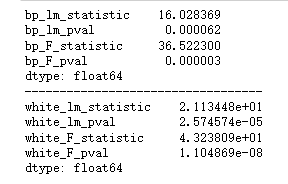




主观判断就是x小的时候方差小，而x大的时候方差会变大

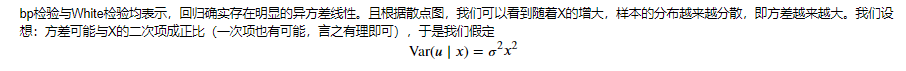
（2）定量BP、white检验





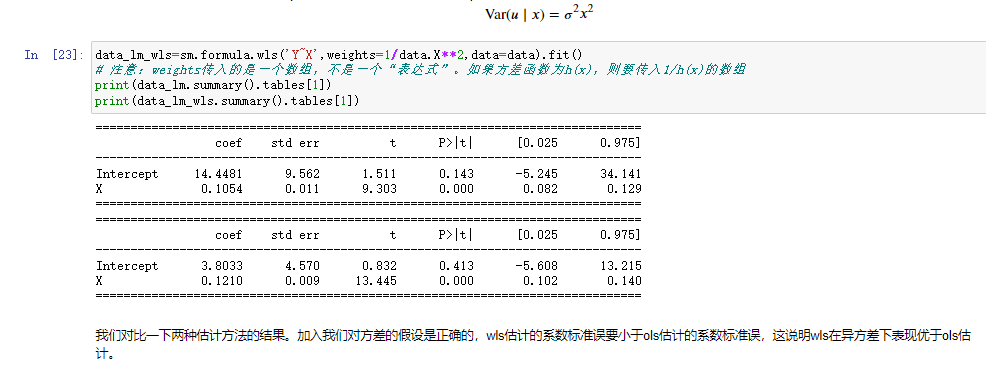
可见p值太小，都会拒绝同方差假设

实践2：如何假设异方差的hx的形式，以使得可以使用WLS



这个猜想好主观，就是假设hx=x2

实践3：就是将权重hx代入

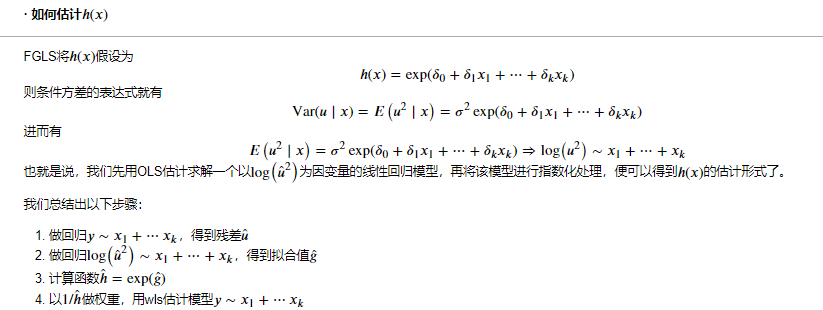


#### 6.3.2 可行的广义最小二乘法

思想：

WLS估计法要求我们知晓ℎ(𝑥)的具体形式，而在大多数情况下，ℎ(𝑥)的形式是难以通过观察得出的，这个时候我们就需要使用一种方法估计出ℎ(𝑥)的形式，再使用wls估计法求解模型。这种方法又称为FGLS。

流程：



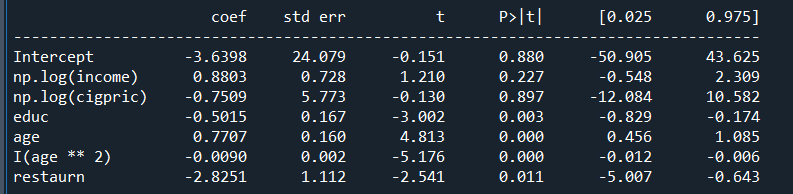
实践：

1. 假设同方差满足的时候，结果为：

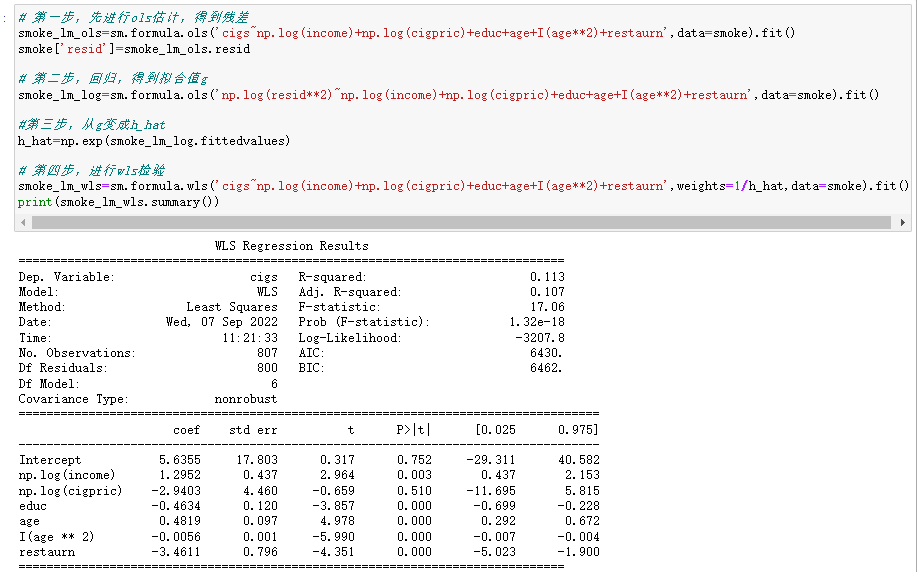
#假设同方差的时候，看估计的标准误

smoke\_lm\_ols=sm.formula.ols('cigs~np.log(income)+np.log(cigpric)+educ+age+I(age\*\*2)+restaurn',data=smoke).fit()

print(smoke\_lm\_ols.summary())



2.使用FGLS



对比可见系数的标准误都变小了

## 7. 分类分析

推断的时候注重可解释性，因此用的还是sklearn.linear\_model中的逻辑回归（是一个分类模型）等模型；如果关注的是分类的精度，就使用xgb，lgb以及NN等

### 7.1 分类问题的类别

分类问题按照被分类对象的类别划分，总共可以分为三类：二分类问题、无序多分类问题、有序多分类问题。这个分类很是新鲜，因为没有关注过无序多分类和有序多分类的问题

1二分类问题可用二分类Logistic回归模型与Probit模型（没有见过）；

2因变量类别多于两个，且类别间没有大小顺序的分类问题，被称为无序多分类问题，可用多类别logistic回归模型建模（其实也就是softmax）。

3有序多分类问题也称为定序问题，在问卷调查中常常出现，如我们要求消费者对产品进行满意度评价，评价结果包括不满意、一般、满意三个选项，我们可以将之赋值为1/2/3。这三个取值是存在先后大小之分的（所以这个时候整数编码，是有实际意义的），这是与无序多分类问题最大的不同。对于这类问题，我们可用有序Logistic回归模型

1. 二分类问题——二分类Logistic回归、Probit回归
2. 无序多分类问题——多分类Logistic回归
3. 有序多分类问题——有序Logistic回归（这个能不能理解为，输出连续的因变量，转换成概率之后，然后按照一定的阈值进行划分，例如0.33和0.66这种，从而实现1/2/3的有序分类）

### 7.2 二分类问题

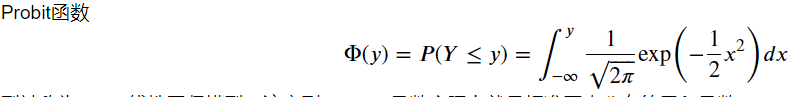
以下是我自己的理解：

分类问题其实也就是将回归问题的连续输出，转换成对应类别的概率

线性概率模型(LPM)不是很出名，因为都得用sigmoid处理成归一化的概率（其实sigmoid函数包括logistic函数，但是机器学习中一般还是称为sigmoid函数更为常见，logistic函数是logit函数的反函数，logit是将0-1之间的概率转换成-inf到inf之间，而logistic是将-inf到inf的连续输出转换到0-1之间的概率输出）

#### 7.2.1****概率的映射****

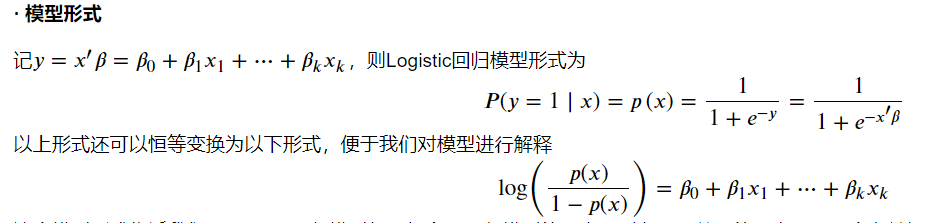




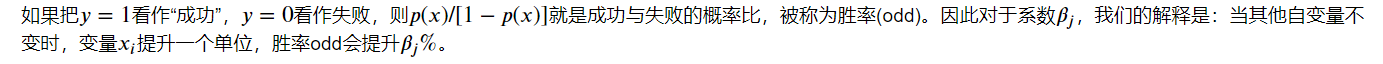
没见过probit函数，好像就是N（0,1）的cdf

好像这两个函数都是S型函数

#### 7.2.2 logistic模型

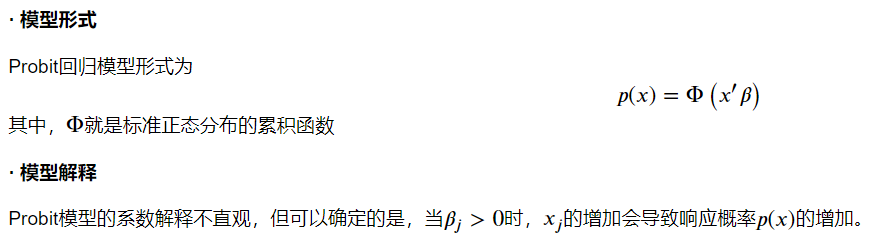


左边其实就是logit函数，也称之为对数几率。



使用sm.formula.logit实现拟合

#### 7.2.3 probit模型（不是很常见）



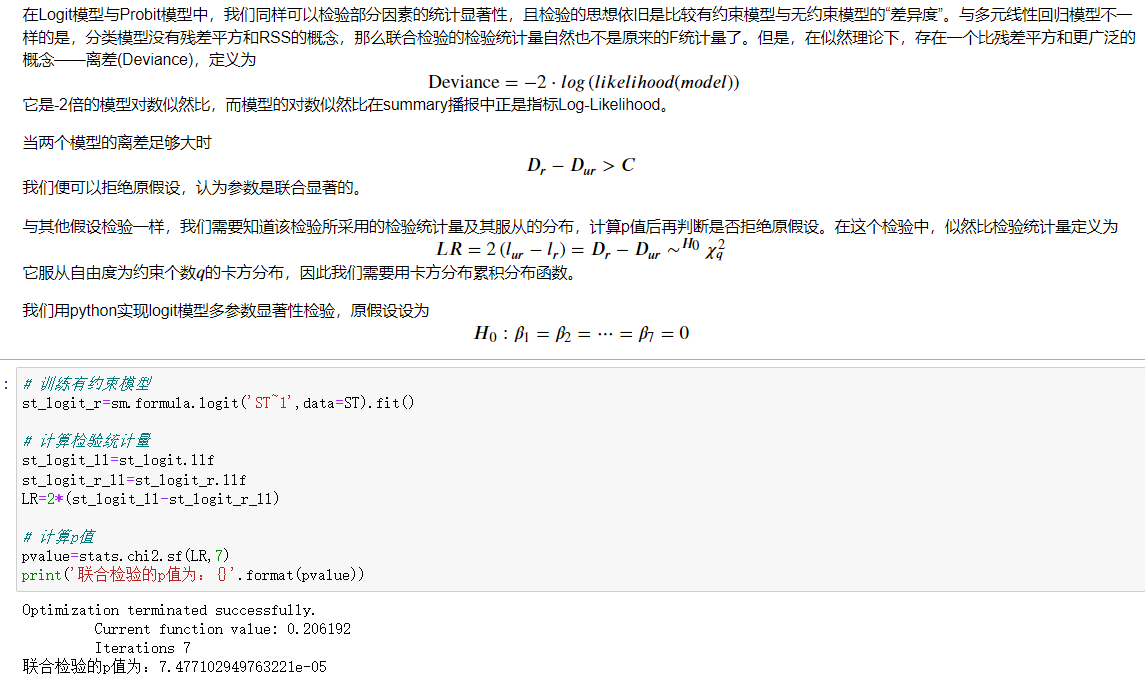
使用sm.formula.probit实现参数的拟合

#### 7.2.4 模型推断

**单参数显著性检验**

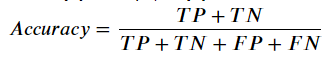


**多参数联合显著性检验**



#### 7.2.5 模型预测（其实sklearn包中有非常好用的函数）

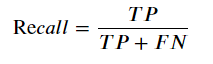
精度：



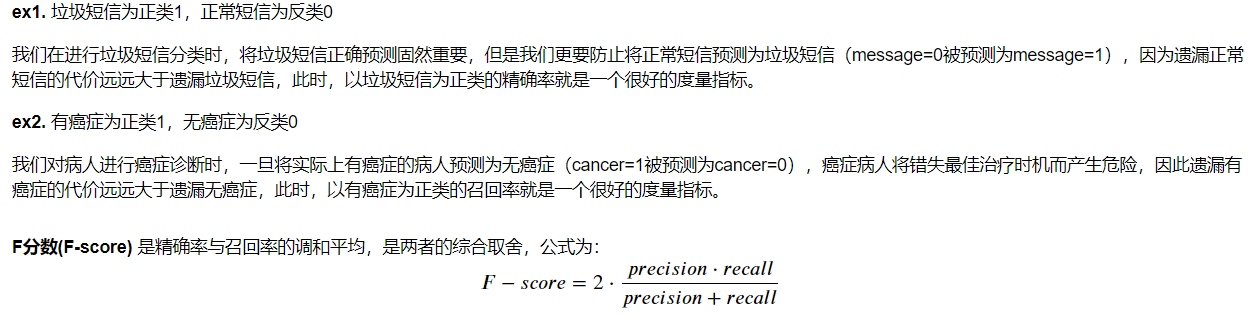
精确率：



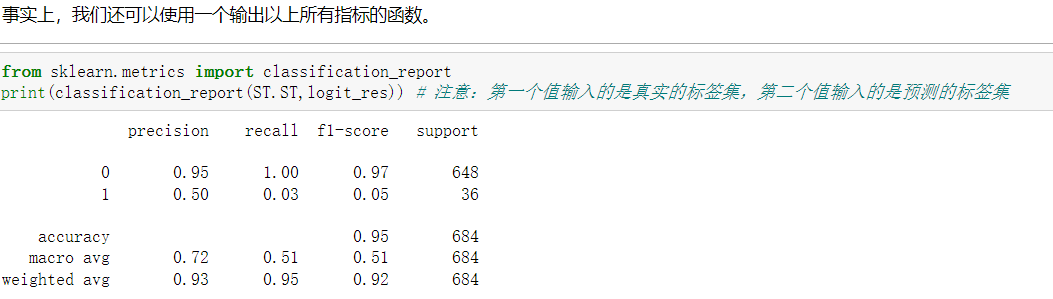
召回率：



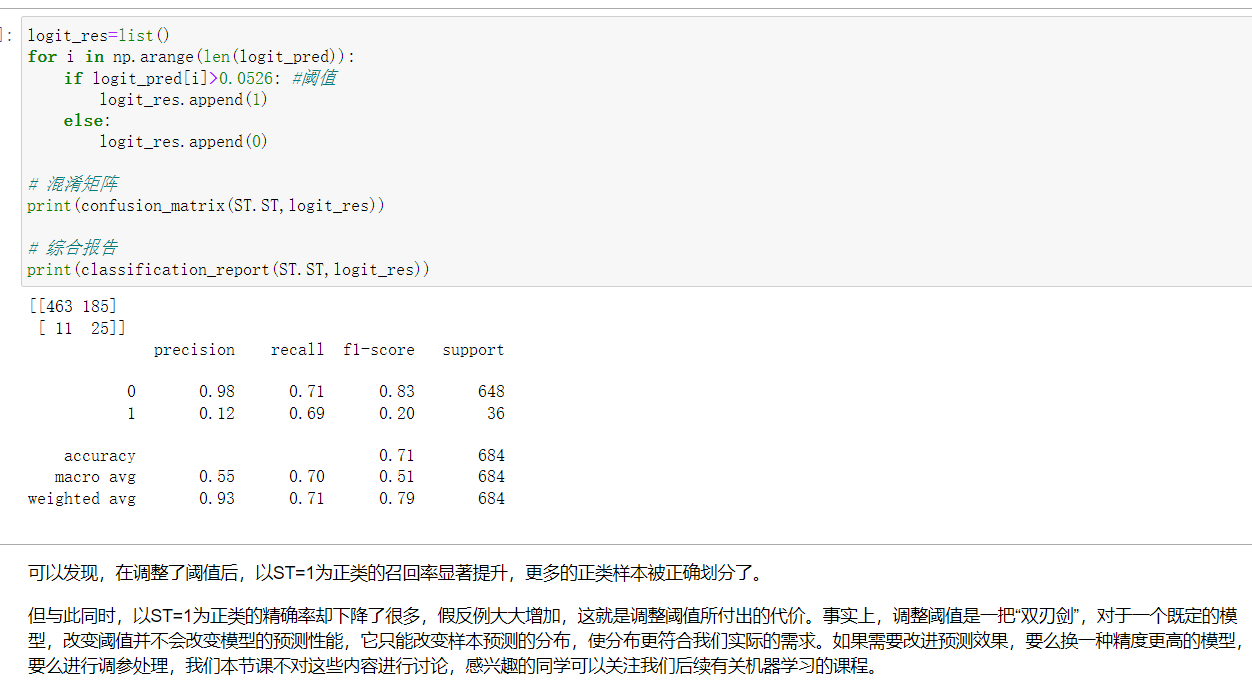
关于这两个指标的例子比较好理解：



使用classification\_report得到所有的指标



#### 7.2.6 不平衡分类问题改变阈值，本质上并不能实现更好的预测效果



### 7.3 无序多分类问题

sklearn也具有做Logistic回归的功能，但它注重预测结果输出，而在统计推断结果的可视化上，sklearn做的远远不如statsmodels。因此本章节我们不介绍多分类Logistic回归的统计推断，只学习如何预测。

不关注显著性检验等等，只关心预测结果

这儿其实就是OvO，OvR等等将二分类模型推广到多分类模型的方法