**OLS估计量（模型正确设定）的无偏性、一致性和渐近正态性**

代码

import numpy as np

import statsmodels.formula.api as smf

from scipy import stats

import statsmodels.api as sm

#OLS估计量（模型正确设定）的无偏性、一致性和渐近正态性

size = 1000

x1 = np.random.exponential(200, size)

x2 = np.random.normal(4, 30, size)

u = np.random.uniform(-1, 1, size)

X = np.column\_stack((x1, x2))

X = sm.add\_constant(X)

e = u\*(sqrt(x1) + 0.05\*x2)

beta = np.array([3, 2, 4])

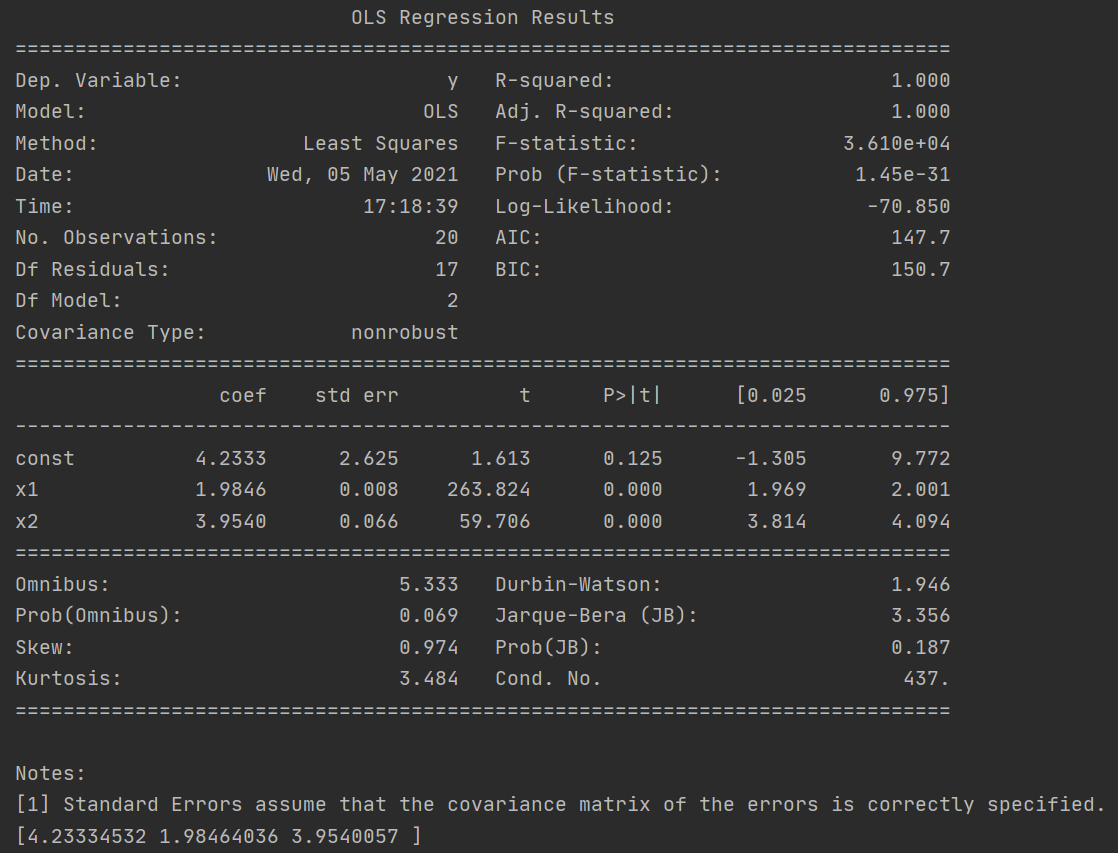
y = np.dot(X, beta) + e

model = sm.OLS(y, X).fit()

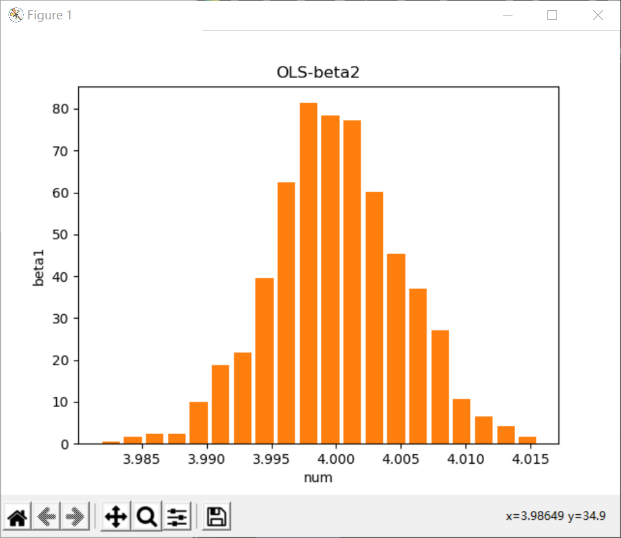
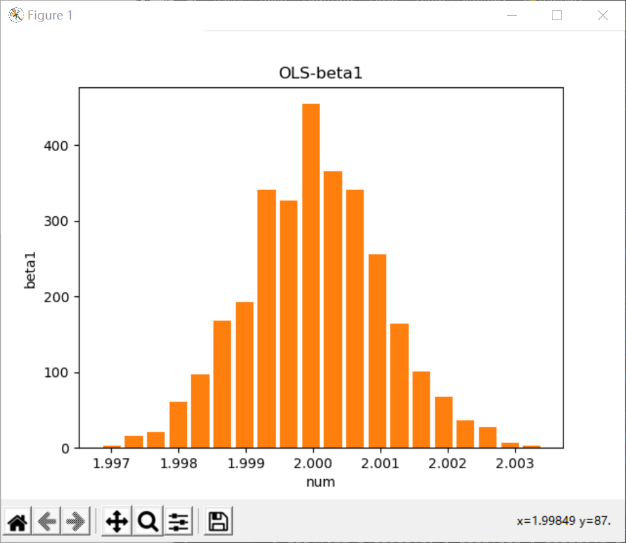
print(model.summary())

print(model.params)

**样本容量为20**



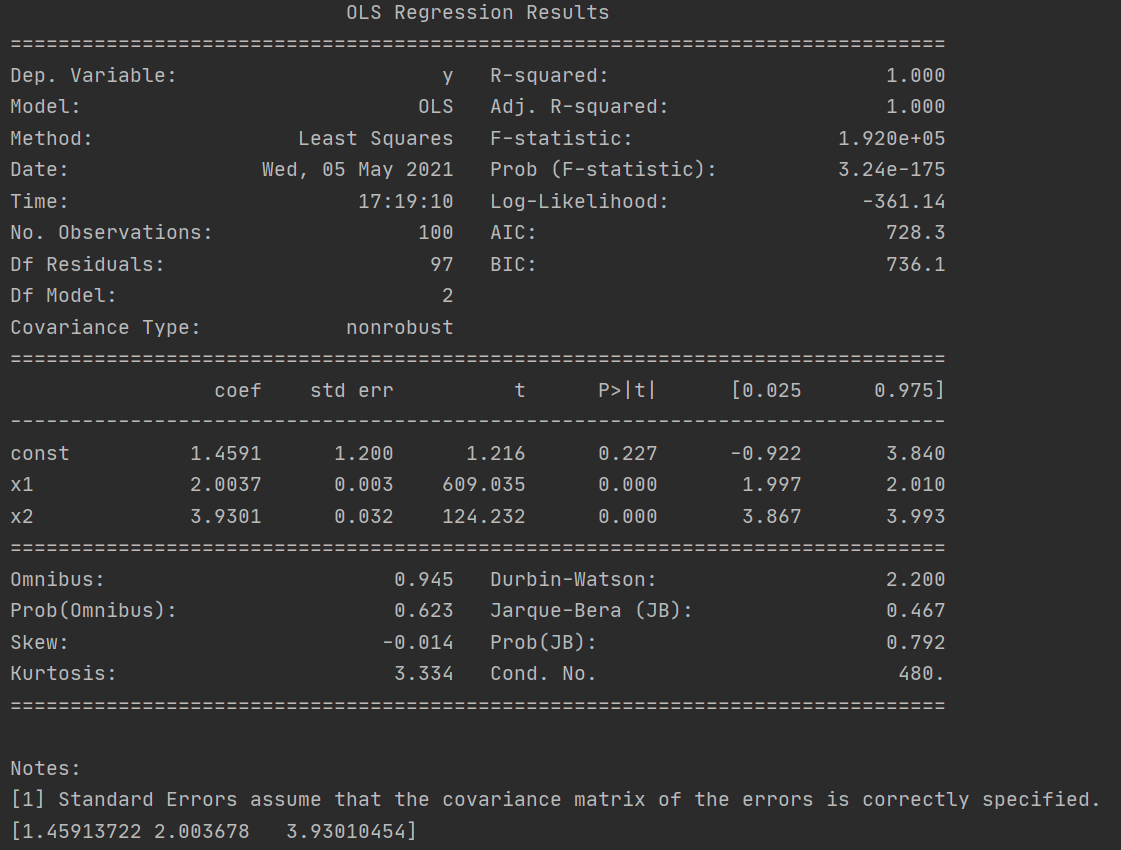
结果显示x1和x2两个变量均显著，但常数项不显著。检验整个方程显著性的F统计量达到3.610e+04，p值为1.45e-31，说明此回归方程高度显著。



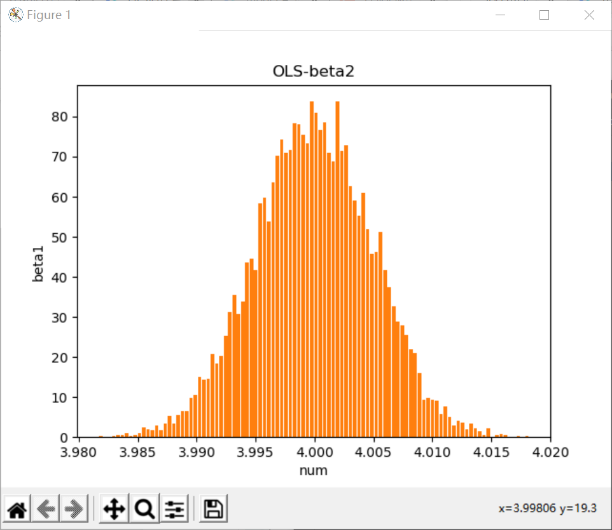
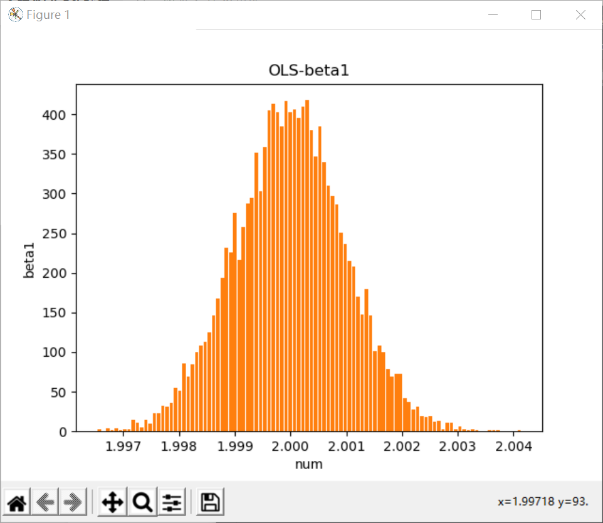
重复抽样3000次后beta1与beta2的分布图，beta1估计偏差为

0.000078，方差为0.00369，Jarque-Bera 统计量的值为9.34，β2估计偏差为0.000567，方差为0.0064，Jarque-Bera 统计量的值为18.88

**样本容量为100**



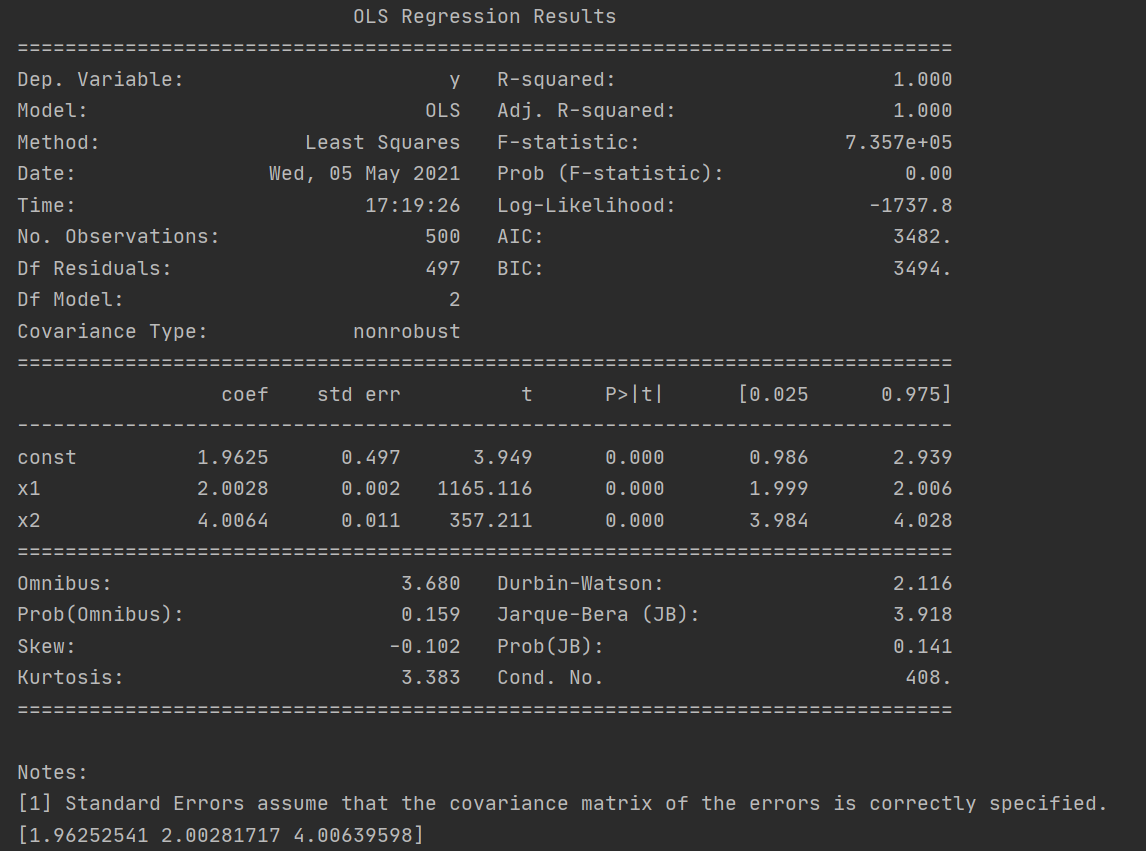
结果显示x1和x2两个变量均显著，但常数项不显著。检验整个方程显著性的F统计量达到1.920e+05，p值为3.24e-175，说明此回归方程高度显著。



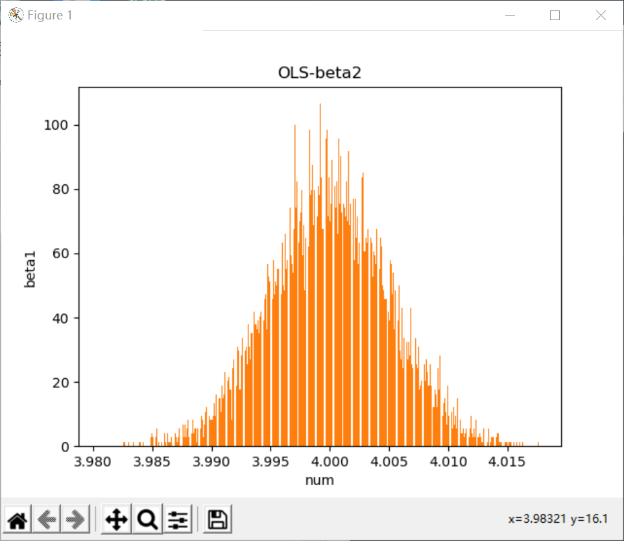
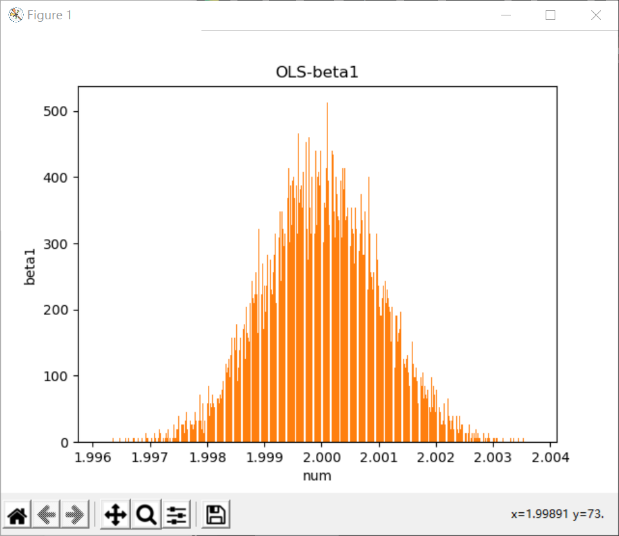
重复抽样3000次后beta1与beta2的分布图，beta1估计偏差为

0.000023，方差为0.0011，Jarque-Bera 统计量的值为3.445，β2估计偏差为0.000345，方差为0.0023，Jarque-Bera 统计量的值为3.084

**样本容量为500**



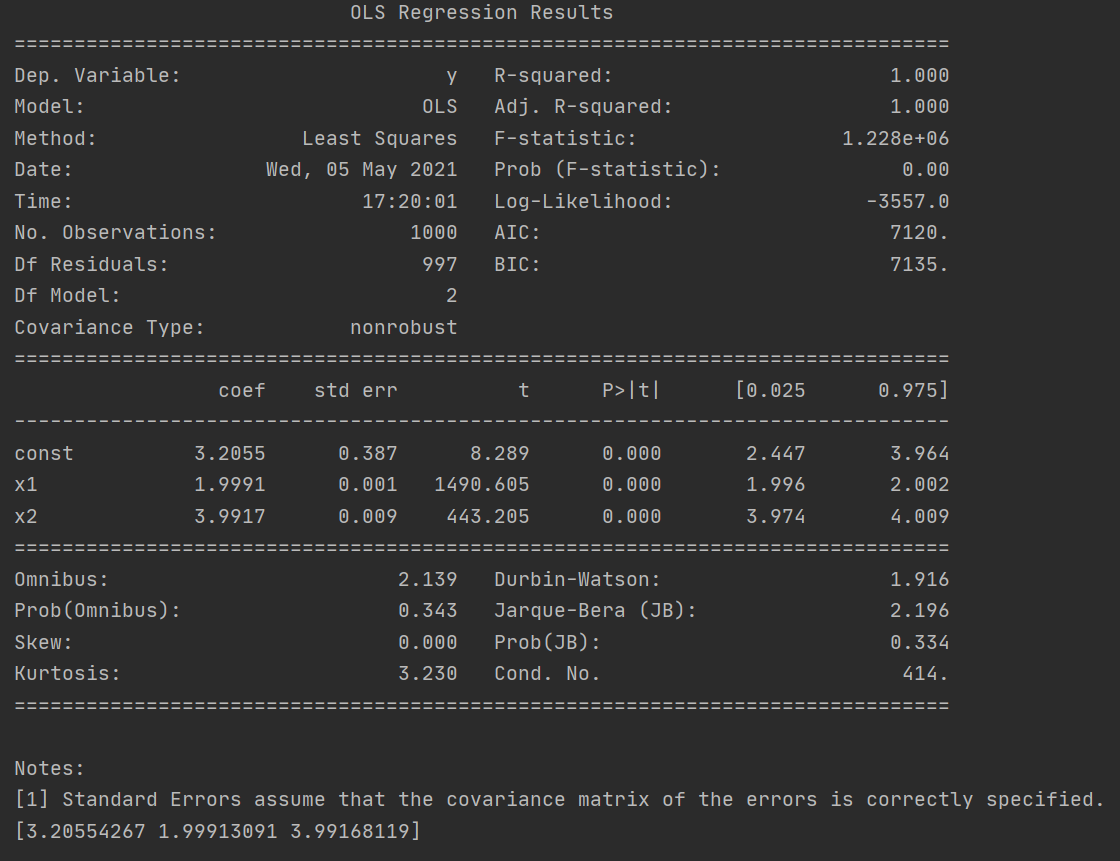
结果显示x1和x2两个变量均显著，但常数项不显著。检验整个方程显著性的F统计量达到7.357e+05，p值为0.0000，说明此回归方程高度显著。



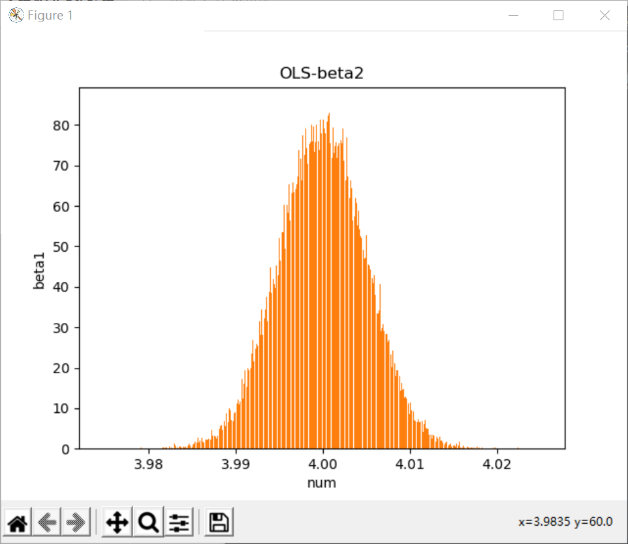
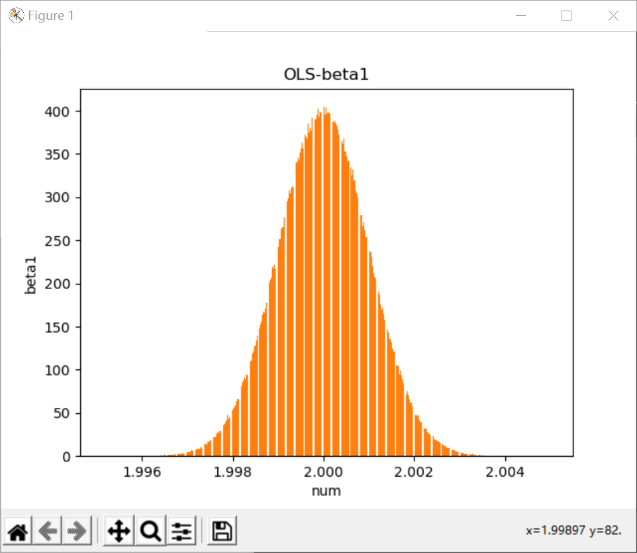
重复抽样3000次后beta1与beta2的分布图，beta1估计偏差为

0.000012，方差为0.00079，Jarque-Bera 统计量的值为0.798，β2估计偏差为0.000142，方差为0.0086，Jarque-Bera 统计量的值为3.039

**样本容量为1000**



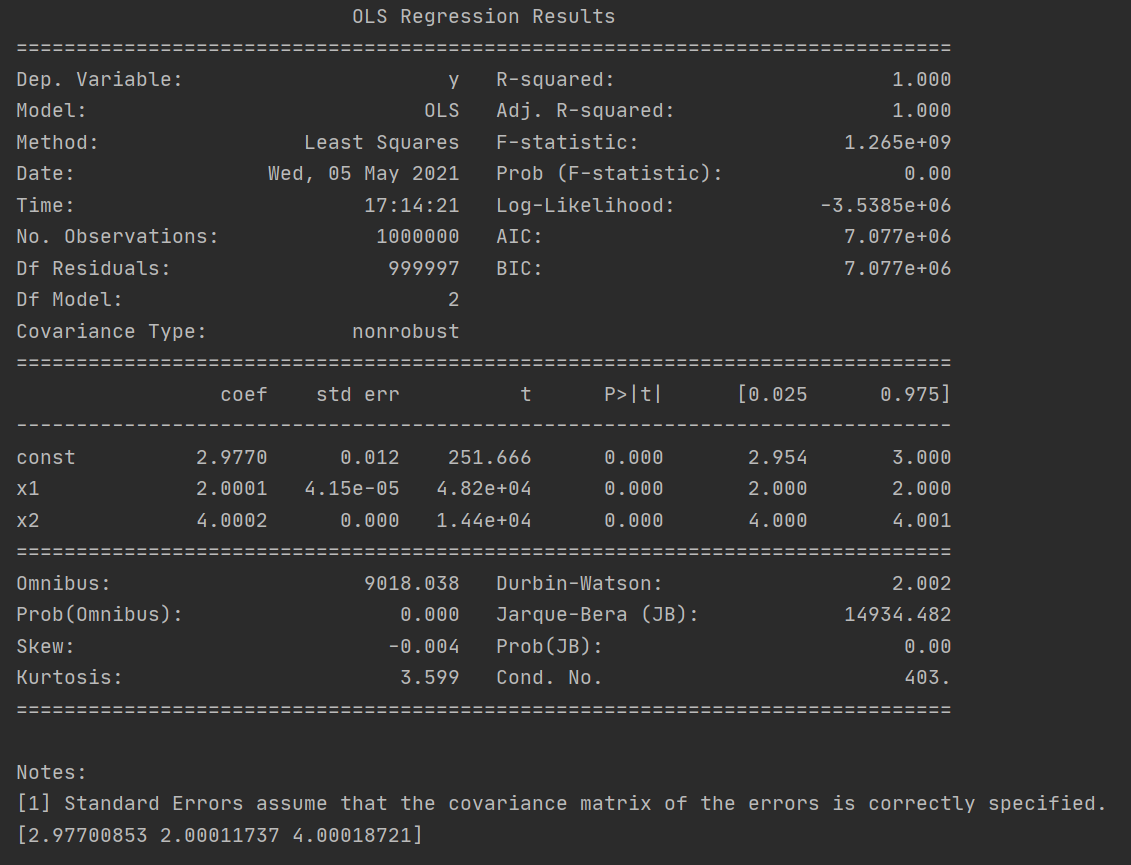
结果显示x1和x2两个变量和常数项都显著。检验整个方程显著性的F统计量达到1.228e+06，p值为0.0000，说明此回归方程高度显著。



重复抽样3000次后beta1与beta2的分布图，beta1估计偏差为

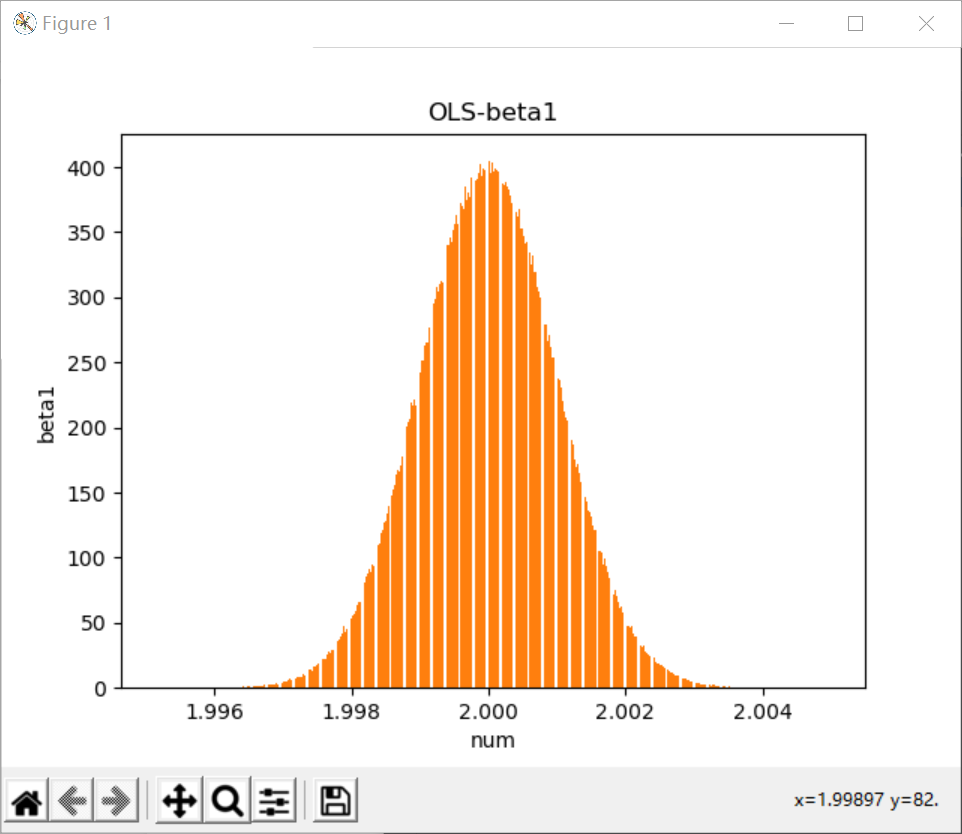
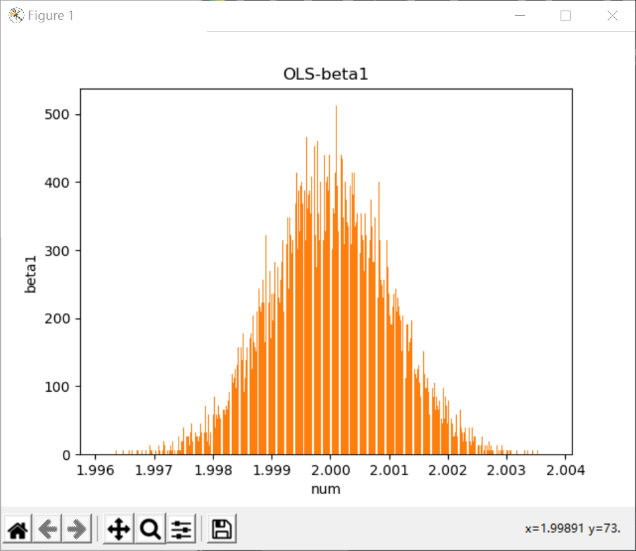
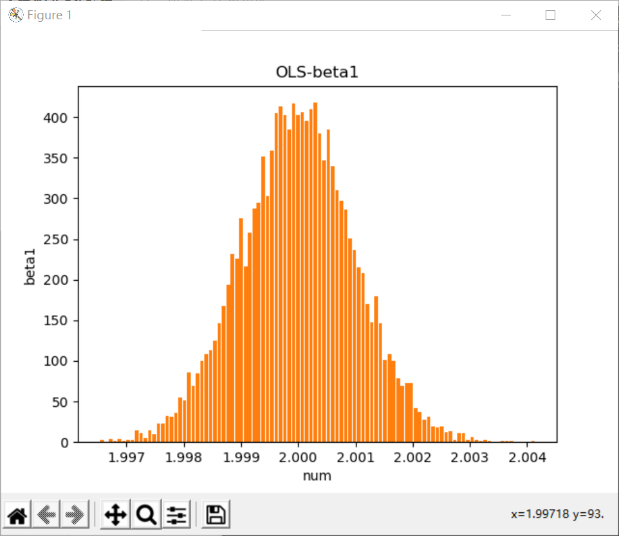
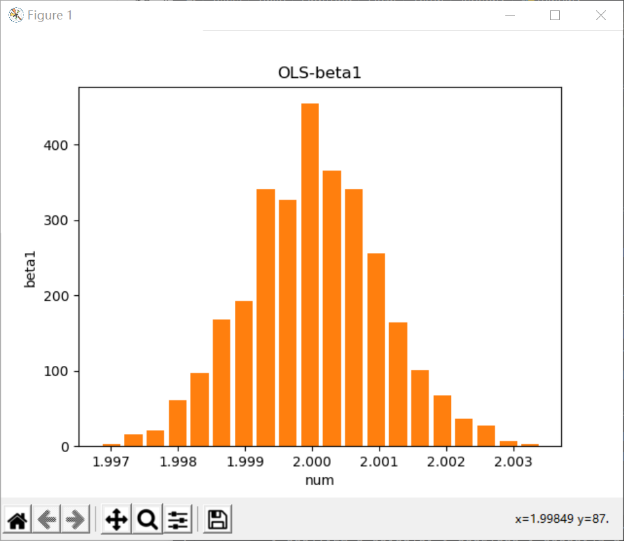
0.000009，方差为0.00012，Jarque-Bera 统计量的值为0.564，β2估计偏差为0.000122，方差为0.0034，Jarque-Bera 统计量的值为3.998

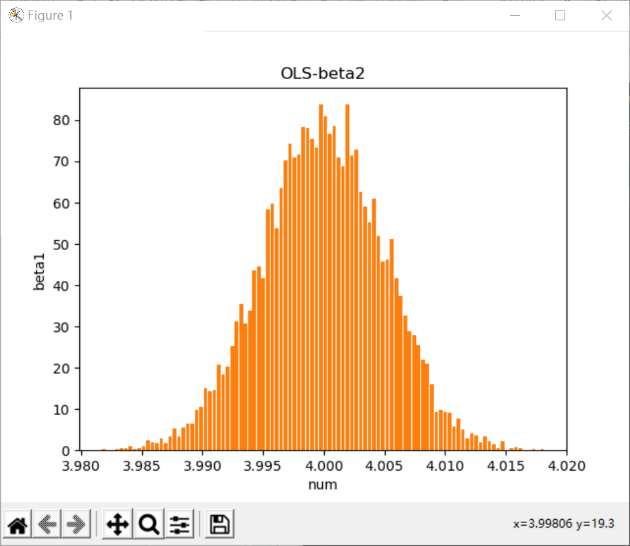
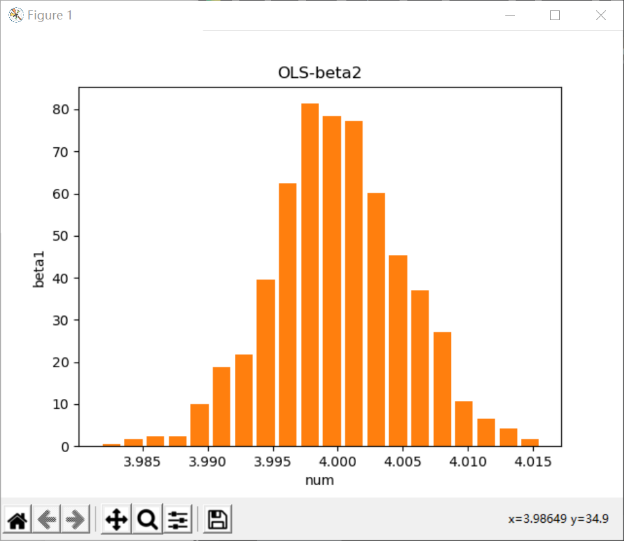
**样本容量为1000000**

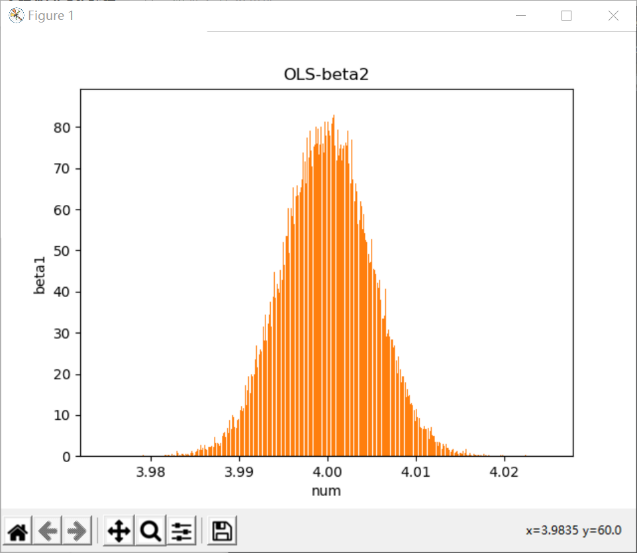
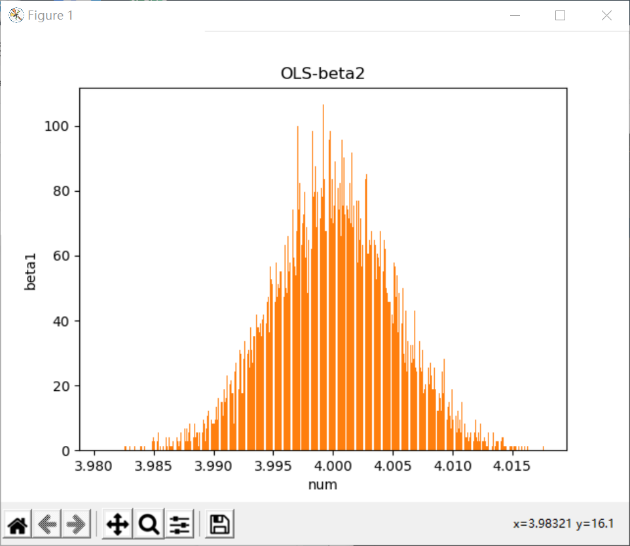


结果显示x1和x2两个变量和常数项均显著。检验整个方程显著性的F统计量达到1.265e+09，p值为0.0000，说明此回归方程高度显著。









**由于随着样本容量不断增大，OLS估计系数越来越接近真实值，系数beta1和beta2估计量的偏差越来越小，参数估计量的平均值与真实参数相吻合，体现无偏性。参数估计量依概率收敛到真实参数，系数beta1和beta2估计量的方差越来越小，体现一致性。与此同时，重复抽样3000次所得的两个参数估计量的Jarque-Bera统计量由原来的远大于零逐渐向零值靠近，说明参数估计量具有渐近正态性。综上所述，当扰动项存在异方差时， OLS估计量（模型正确设定）具有无偏性、一致性、渐近正态性。**

**在不知道异方差真实形式的情况下，用White检验异方差的形式，并根据检验结果构造FGLS估计过程，考察估计结果的统计性质，并将其与OLS的结果进行比较**

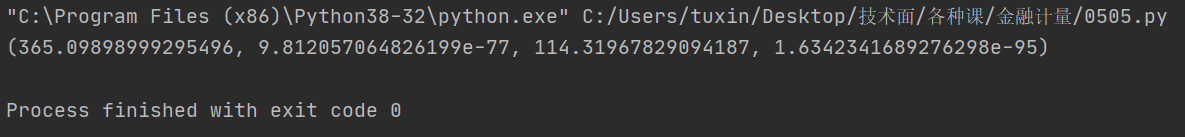
**White检验**

**样本容量为1000**

print(sm.stats.diagnostic.het\_white(model.resid, exog = model.model.exog))

White检验结果

(365.09898999295496, 9.812057064826199e-77, 114.31967829094187, 1.6342341689276298e-95)



可以看出P值为9.812057064826199e-77为0，故强烈拒绝同方差的原假设，认为存在异方差

**样本容量为20**

可以看出P值为1.4633e-34为0，故强烈拒绝同方差的原假设，认为存在异方差

**样本容量为100**

可以看出P值为7.35709e-53为0，故强烈拒绝同方差的原假设，认为存在异方差

**样本容量为500**

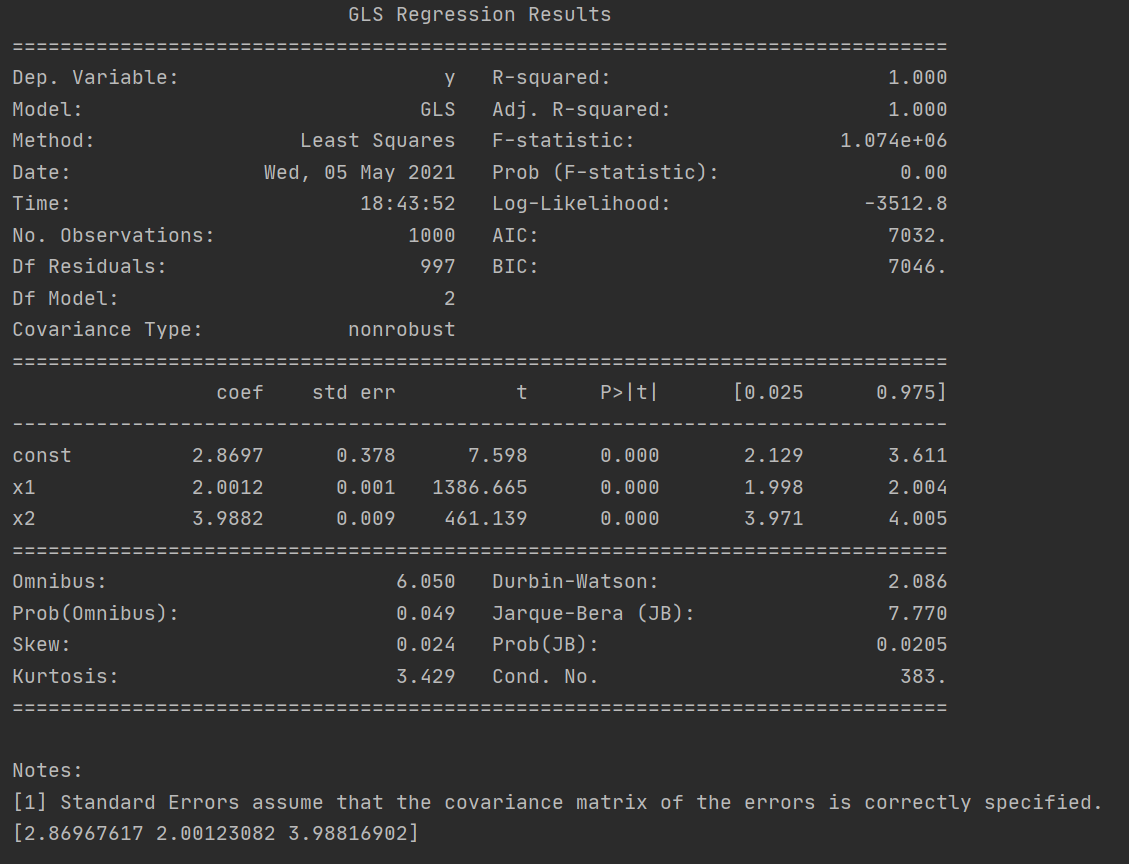
可以看出P值为1.331209e-59为0，故强烈拒绝同方差的原假设，认为存在异方差

**FGLS估计过程（以样本容量1000为例）**

model2 = sm.GLS(y, X).fit()

print(model2.summary())

print(model2.params)



结果显示，p值等于0.0000，说明拒绝同方差的原假设，存在异方差。然后进行FWLS，计算残差并记为e1,生成平方e2，得到扰动项方差的估计值，为保证方差始终为正，再取对数得到lne2。分别用lne2和x1、x2进行OLS回归，同时考虑是否有常数项，得到的最好结果显示去掉常数项和x1进行辅助回归得到的R^{2}为0.4858，说明解释变量x1可以解释lne2 48.58%的变动，残差平方的变动与x1在一定程度上相关。

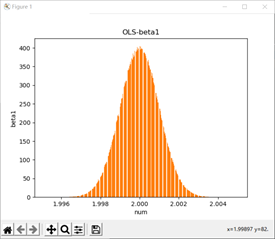
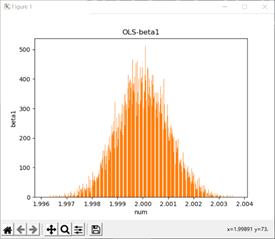
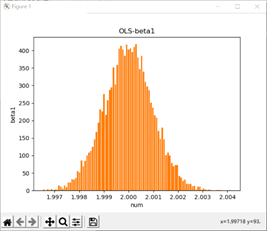
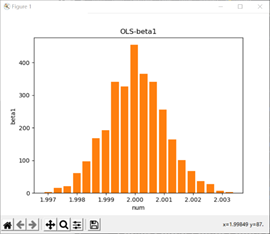
**与大样本OLS比较**

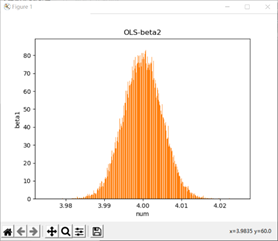
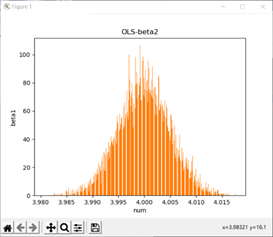
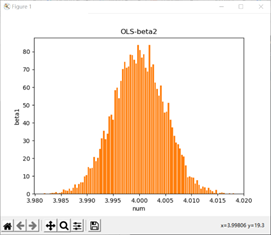
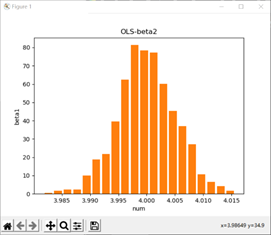
当样本容量增大时，FWLS方法系数估计的偏差值和稳健标准误都下降，因为FWLS的优点主要体现在大样本理论中。其次，当样本容量取值相同时，大样本OLS方法得到系数估计的偏差好于FWLS方法，而FWLS的稳健标准误则低于OLS，可能的原因在于：

① FWLS在大样本下比OLS更有效率；② FWLS中对条件方差函数具体形式的设定不一定是最准确的，因此导致结果的偏差。

**在不知道自相关真实形式的情况下，用DW, Ljung-Box-Q、BG方法，检验自相关的形式，并根据检验结果构造PW估计过程，考察估计结果的统计性质，并将其与OLS的结果进行比较**

首先用OLS进行分析





和扰动项为异方差的实验相比，自相关实验中，样本容量较小时模型已经拟合得非常好，随着样本容量的不断增大，系数beta1和beta2估计量的偏差呈减小的趋势，参数估计量的平均值与真实参数相吻合，体现无偏性。参数估计量依概率收敛到真实参数，系数beta1和beta2估计量的方差越来越小，体现一致性。重复抽样3000次所得的两个参数估计量的Jarque-Bera统计量由原来的远大于零逐渐向零值靠近，说明参数估计量具有渐近正态性。综上所述，当扰动项为自相关时，大样本OLS估计量具有无偏性、一致性和渐近正态性。

**DW检验**

from statsmodels.stats.stattools import durbin\_watson

print(durbin\_watson(model.resid))

使用**Durbin Watson**统计量检查残差项的序列相关性的公式如下: DW值可以在0到4之间变化。它越接近值2，则没有明显的序列相关性。接近0时，存在正序列相关，而接近4时，则具有负序列相关。

样本容量为20

检验结果1.1999879236060504离2较远离0较近，可大致判断存在自相关性。

样本容量为100

检验结果1.09离2较远离0较近，可大致判断存在自相关性。

样本容量为500

检验结果0.785456离2较远离0较近，可大致判断存在自相关性。

样本容量为1000

检验结果0.4536离2较远离0较近，可大致判断存在自相关性。

**BG检验**

进行BG检验， p值为0.00075，在1%的显著性水平上拒绝原假设，认为存在自相关。

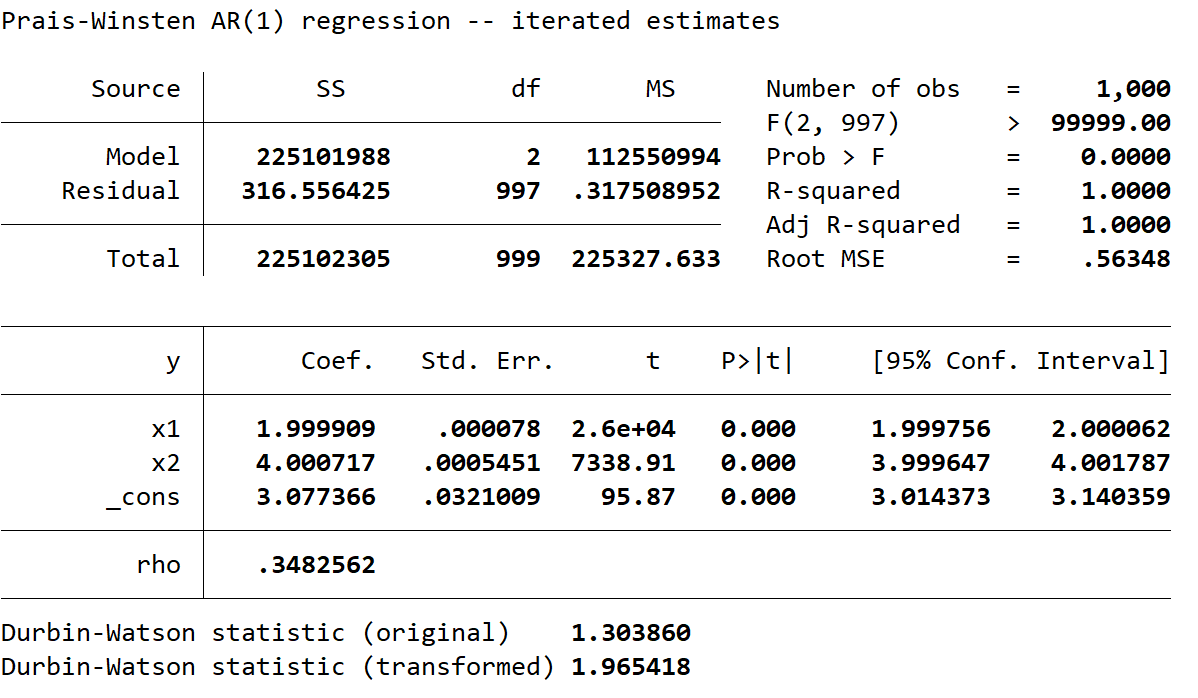
**Ljung-Box-Q检验**

from statsmodels.stats.diagnostic import acorr\_ljungbox

print(acorr\_ljungbox(model.resid, lags = 3, boxpierce=True))

进行Q检验。p值为0.0057，在1%的显著性水平上拒绝原假设，认为存在自相关。

**PW估计法（以样本容量1000为例）**



**PW与OLS相比较**

PW估计法表现得没有大样本OLS稳健，样本容量增大时，其系数估计的偏差反而也增大。

如果自相关系数不满足严格外生性，仅满足前定解释变量，则FGLS可能不一致，因此FGLS的适用条件更苛刻。