

目录

关于 VAR 宏观经济模型的初步研究..... 2

1, 数据获取 ..... 2

2, 选择模型 ..... 2

    VAR 模型又称向量自回归模型 ..... 2

    Var 模型的优点 ..... 3

3. 代码编写 ..... 3

# 关于 VAR 宏观经济模型的初步研究

## 1, 数据获取

选择了 11 个左右的协变量数据源:

Index(['美国国内公司市值', '人均 GPD', '劳动力人口占比', '年度仇恨犯罪数', '年度贫困率', '年均失业率', '年人均收入中位数', '通胀率', '年平均基金利率', '年出口额', '年进口额']

因变量: 人均 GPD, 年人均收入中位数

协变量: '美国国内公司市值', '劳动力人口占比', '年度仇恨犯罪数', '年度贫困率', '年均失业率', '通胀率', '年平均基金利率', '年出口额', '年进口额'

截取自 2004-2020 的数据, 均来源于 US 官方网站  
从各指标 csv 中提取并整合到一个 csv 表中

	🇺🇸 年份	💎 美国国内...	💎 人均GPD	💎 劳动力人...	💎 年度仇恨...	💎 年度贫困率	💎 年均失业率	💎 年人均收...	💎 通胀率	💎 年平均基...	💎 年出口额	💎 年进口额
1	2004	16300000000...	41724.63	0.740	7649	12.7	5.550	34430	0.012	0.013	823584	1488348
2	2005	17000000000...	44123.41	0.740	7163	13.3	5.575	35070	0.018	0.032	913016	1695821
3	2006	19600000000...	46302.00	0.741	7722	13.3	4.675	36140	0.024	0.050	1040906	1878194
4	2007	19900000000...	48050.22	0.738	7624	13.0	4.650	36370	0.021	0.050	1165150	1986347
5	2008	11600000000...	48570.05	0.738	7783	13.2	5.000	34930	0.013	0.019	1308794	2141287
6	2009	15100000000...	47194.94	0.731	6604	14.3	9.675	34570	0.001	0.002	1070330	1580025
7	2010	17300000000...	48650.64	0.724	6628	15.3	9.167	34160	0.016	0.002	1290278	1938952
8	2011	15600000000...	50065.97	0.719	6222	15.9	8.475	33650	0.015	0.001	1498886	2239885
9	2012	18700000000...	51784.42	0.718	5796	15.9	7.900	33510	0.021	0.001	1562630	2303749
10	2013	24000000000...	53291.13	0.716	5928	15.8	7.142	33950	0.032	0.001	1593708	2294247
11	2014	26300000000...	55123.85	0.715	5479	15.5	6.267	34780	0.016	0.001	1635563	2385480
12	2015	25100000000...	56762.73	0.714	5850	14.7	5.608	36610	-0.004	0.001	1511381	2273249
13	2016	27400000000...	57866.74	0.717	6121	14.0	4.883	37310	0.038	0.004	1457392	2207194
14	2017	32100000000...	59907.75	0.721	7175	13.4	4.242	37690	0.029	0.010	1557003	2356345
15	2018	30400000000...	62823.31	0.724	7120	13.1	3.700	38960	0.032	0.018	1676913	2555662
16	2019	33900000000...	65120.39	0.729	7314	12.3	3.800	40980	0.034	0.022	1655098	2512358
17	2020	40700000000...	63528.63	0.717	8263	11.9	7.533	40420	0.027	0.004	1433852	2346727
18	2021		70219.47	0.721		12.8						
19	2022		76329.58			12.6						

## 2, 选择模型

1. Jamovi, 新兴的统计分析软件, 但在尝试后发现 jamovi 不支持时间序列分析, 只支持分组的数据分析, 因此放弃该软件选择, 转而使用 python+jupyter notebook 的形式使用其他金融数学模型

结论: jamovi 可以直观地分析数据样本, 但不便于预测

2. 在决定使用时间序列模型之后, 根据[相关文章](#)和适用领域的推荐, 决定采用 VAR 模型。

VAR 模型又称向量自回归模型, 是一种常用的计量经济模型, 该模型扩充了只用一个变量的自回归模型, 是采用多方程联立的形式, 不以经济理论为基础, 在模型的每一个方程中, 内生变量对模型的全部内生变量的滞后期进行回归, 从而估计全部内生变量的动态关系, 并进行预测。该模型使得时间序列分析从单一时间序列拓展到了多元时间序列, 被广泛的应用在金融和经济等多个领

域。可以说，只要问题涉及多变量，时间序列数据，都有可能使用 VAR 模型。

以两个变量  $y_{1t}, y_{2t}$  滞后 1 期的 VAR 模型为例，VAR 模型可表达为：

$$y_{1t} = c_1 + \pi_{11,1}y_{1,t-1} + \pi_{12,1}y_{2,t-1} + u_{1t}$$

$$y_{2t} = c_2 + \pi_{21,1}y_{1,t-1} + \pi_{22,1}y_{2,t-1} + u_{2t}$$

写成矩阵形式如下所示：

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} \pi_{11,1} & \pi_{12,1} \\ \pi_{21,1} & \pi_{22,1} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}$$

设

$$Y_t = \begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix}, c = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}, \Pi_1 = \begin{bmatrix} \pi_{11,1} & \pi_{12,1} \\ \pi_{21,1} & \pi_{22,1} \end{bmatrix}, Y_{t-1} = \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix}, u_t = \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}$$

则 矩阵形式可以表示为：

$$Y_t = c + \Pi_1 Y_{t-1} + u_t$$

Var 模型的优点在于：

- 第一，不以严格的经济理论为依据，而是让数据关系说明一切；
- 第二，解释变量中不包括任何当期变量，只要样本足够大，就不存在因参数过多产生模型不可识别的问题；
- 第三，无需事先区分变量的外生性和内生性

结论：VAR 模型常被用于宏观经济领域的多变量预测，和目前的需求具有高度相关性，准备建模。

## 3. 代码编写

### ● 相关库，数据导入（良好运行）

```
import pandas as pd
import numpy as np
from statsmodels.tsa.vector_ar.var_model import VAR
from statsmodels.tsa.stattools import adfuller
import matplotlib.pyplot as plt

# 加载数据并设置时间索引
df = pd.read_csv('input.csv')
#for i in ['劳动力人口占比', '年度仇恨犯罪数', '年度贫困率', '年均失业率', '年平均基金利率', '年出口额', '年进口额']:
#    df.drop(i, axis='columns', inplace=True)

df['Year'] = pd.to_datetime(df['Year'], format='%Y')
df.set_index('Year', inplace=True)
```

- 数据处理（意义存疑，理论上数据平稳性应该够高，但最后没有成功，不得不加上作为尝试）

```
# 检查平稳性并进行差分
def check_stationarity(ts, alpha=0.05, name='', verbose=False):
    result = adfuller(ts.dropna(), autolag='AIC')
    output = {'test_statistic': round(result[0], 4), 'pvalue': round(result[1], 4), 'n_lags': result[2], 'n_obs': result[3]}
    p_value = output['pvalue']
    if verbose:
        print(f'ADF Statistic: {output["test_statistic"]} for {name}')
        print(f'p-value: {output["pvalue"]} for {name}')
    return p_value <= alpha

stationary_data = pd.DataFrame()
for column in df.columns:
    if not check_stationarity(df[column], name=column, verbose=True): # 打开详细输出来调试
        stationary_data[column] = df[column].diff().dropna()
    else:
        stationary_data[column] = df[column]
stationary_data.dropna(inplace=True) # 删除因差分造成的任何NaN值
```

- 创建模型选择参数（重大错误，最大延迟数长期有错误，多次研究后发现数据数量与变量数不符，需

```
# 使用 VAR 模型自动选择滞后数
model = VAR(stationary_data)
#max_lag = min(15, len(stationary_data) // (len(df.columns) * 3)) # 计算更合理的滞后数
max_lag = 2
selected_model = model.fit(maxlags=max_lag, ic='aic')
print(f'Optimal number of lags: {selected_model.k_ar}')
print(selected_model.summary())
```

返回第一步收集数据重新整理数据，预计需要收集 1980s 的数据）

- 预测（未能进行到这一步）

```
# 冲击响应分析
try:
    irf = selected_model.irf(5)
    irf.plot(orth=True)
    plt.show()
except Exception as e:
    print("Error computing or plotting IRF:", e)

# 预测
lag_order = selected_model.k_ar
if lag_order >= len(stationary_data):
    print("Not enough data points for the number of lags used.")
else:
    forecast_input = stationary_data.values[-lag_order:]
    forecasted_values = selected_model.forecast(forecast_input, steps=5)
    print("Forecasted values:")
    print(forecasted_values)

# 将预测转换为 DataFrame
forecast_df = pd.DataFrame(forecasted_values, index=pd.date_range(start=stationary_data.index[-1], periods=6, freq='A')[1:], columns=s)
print(forecast_df)

# 保存预测结果
forecast_df.to_csv('forecasted_output.csv')
```

● 目前的结果

Correlation matrix of residuals

	美国国内公司市值	人均GDP	劳动力人口占比	年度仇恨犯罪数	年度贫困率	年均失业率	年人均收入中位数	通胀
率	1.000000	-0.206579	-0.341949	0.202862	-0.147022	0.261816	0.219635	0.367850
年平均基金利率	-0.206579	1.000000	0.100289	0.065233	-0.282485	-0.876623	0.594671	0.308875
年出口额	0.100289	0.100289	1.000000	-0.159506	0.356379	0.128987	-0.129368	-0.068035
年进口额	-0.341949	0.065233	-0.159506	1.000000	-0.630374	-0.198469	0.156997	0.436198
美国国内公司市值	0.202862	0.065233	-0.159506	1.000000	-0.630374	-0.198469	0.156997	0.436198
人均GDP	-0.147022	-0.282485	0.356379	-0.630374	1.000000	0.318554	-0.597841	-0.449117
劳动力人口占比	0.261816	-0.876623	0.128987	-0.198469	0.318554	1.000000	-0.386370	-0.281777
年度仇恨犯罪数	0.219635	0.594671	-0.129368	0.156997	-0.597841	-0.386370	1.000000	0.221694
年度贫困率	0.367850	0.308875	-0.068035	0.436198	-0.449117	-0.281777	0.221694	1.000000
年均失业率	-0.198370	0.371320	0.836732	0.120738	-0.072177	-0.157201	0.206744	0.183859
年人均收入中位数	-0.436480	0.671428	0.285502	-0.001152	0.231823	-0.719444	-0.145620	1.000000
通胀率	-0.357196	0.678982	0.145324	0.205811	0.074947	-0.747395	-0.053856	1.000000

Summary of Regression Results

Model:	VAR		
Method:	OLS		
Date:	Thu, 09, May, 2024		
Time:	11:30:44		
No. of Equations:	11.0000	BIC:	107.852
Nobs:	16.0000	HQIC:	107.348
Log likelihood:	-1097.30	FPE:	4.07213e+46
AIC:	107.321	Det(Omega_mle):	2.09027e+46

Results for equation 美国国内公司市值

	coefficient	std. error	t-stat	prob
const	1525000000000.000488	894869636688.309326	1.704	0.088

- 错误总结：VAR 模型框架已经搭建，但实际运行后数据源却不达标，参数与数据意义尚待研究
- 可能的解决方案：
- 1，重新收集数据，继续套入模型
  - 2，极大程度的减少变量，仅一个自变量和一个因变量
  - 3，更换模型