上海理工大学光电信息与计算机工程学院

《智能仿真实验》报告

专　　业　智能科学与技术

姓　　名　高浩琦

学　　号　2035060413

年　　级　2020 级

指导教师　陈玮

成　　绩：

教师签字：

〇、实验内容

1. 以中美两国**1980年至2016年的GDP历史数据**为基础，用**多项式拟合**进行曲线拟合，确定其数学模型，并给出拟合过程及分析。
2. 用最佳的多项式拟合模型进行**最小二乘拟**合，并观察**模型误差。**
3. 用拟合模型**预测 2017年至2020年**中、美GDP，并计算模型的**均方误差MSE和拟合优度R^2。**

一、实验原理

- **协方差矩阵**
  - $corrcoef(X,Y)=\begin{bmatrix} Cov(X,X) & Cov(X,Y) \\ Cov(Y,X) & Cov(Y,Y) \end{bmatrix}$
  - 数字越接近于1 说明相关性越强
- **Feature scaling(特征缩放)--Standardization(标准化)**
  - 对不同特征维度的伸缩变换的目的是使得不同度量之间的特征具有**可比性,**对目标函数的影响体现在**几何分布上，同时不改变原始数据的分布。**
  - 特征标准化使数据中每个特征的值具有**零均值（当减去分子中的均值时）和单位方差。**该方法广泛用于许多机器学习算法（例如，支持向量机、逻辑回归和人工神经网络）中的归一化。

    $$X^*=\frac{X-\bar{X}}{\sigma}$$
- **R-square(R^2)(决定系数):Coefficient of determination****

  R^2 即判定系数，也称为**拟合优度**

  拟合优度越大，自变量对因变量的解释程度越高，自变量引起的变动占总变动的百分比就越高。观察点在回归直线附近越密集。取值范围：0~1.

$$R_2=1-\frac{SS\ res}{SS\ tot}$$

$$R^2=1-\frac{\sum_i(\hat{y}_i-y_i)^2}{\sum_i(y_i-\bar{y})^2}$$

> $SS\ reg=\sum_{i=1}^{n}(\hat{y}-\bar{y})^2$
>
> $SS\ res=\sum_{i=1}^{n}(y_i-\hat{y})^2=e_i^2$
>
> $SS\ tot=\sum_{i=1}^{n}(y_i-\bar{y})^2$
>
> 回归平方和：**SSR(Sum of Squares for regression) = ESS (explained sum of squares)**
>
> 残差平方和：**SSE(Sum of Squares for Error) = RSS(residual sum of squares)**
>
> 总离差平方和：**SST(Sum of Squares for total) = TSS(total sum of squares)**
>
> **SSE+SSR=SST**
>
> **RSS+ESS=TSS**

- **SSE(残差平方和):Sum of Squares for Error**

  $\sum_{i=1}^{n} w_i(y_i - \hat{y})^2 = e_i^2$

  $SS\ res\ = \sum_{i=1}^{n} (y_i - \hat{y})^2 = e_i^2$

  - **越接近0说明越有效**
- **RMSE(均方根误差、标准差)：Root mean squared error**

  or **Root-mean-square deviation** (**RMSD**) 均方根偏差

$$RMSE = \sqrt{MSE} = \sqrt{\frac{SSE}{n}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y_i})^2}$$

每个误差对 RMSD 的影响与**平方误差的大小成正比**;

因此较大的误差对 RMSD 的影响非常大。因此，RMSD 对异常值很敏感。

> **MSE(均方差、方差)**：Mean squared error
>
> 数理统计中均方误差是指**参数估计值与参数值之差平方的期望值**，记为**MSE**。MSE是衡量**"平均误差"**的一种较方便的方法，**MSE可以评价数据的变化程度，MSE的值越小，说明预测模型描述实验数据具有更好的精确度。**预测数据和原始数据对应点误差平方和的均值
>
> $MSE = \frac{SSE}{n}$

- **最小二乘法验证（见word）**

二、实验过程及代码分析

1. 相关性分析

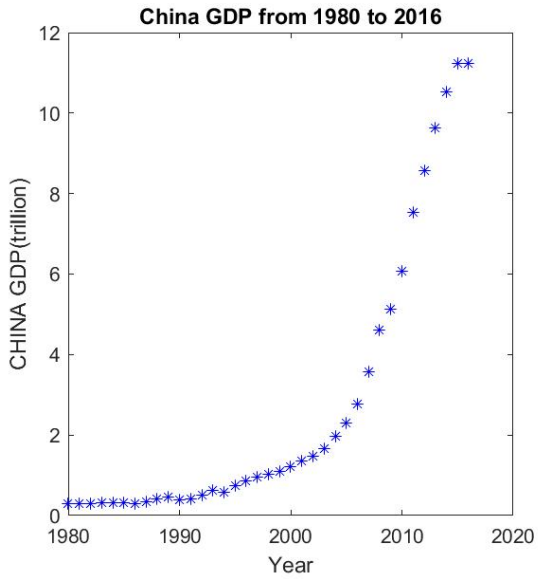
```
CH_corr = corrcoef(y,CH)
US_corr = corrcoef(y,US)

CH_corr =

    1.0000    0.8394
    0.8394    1.0000

US_corr =

    1.0000    0.9933
    0.9933    1.0000
```

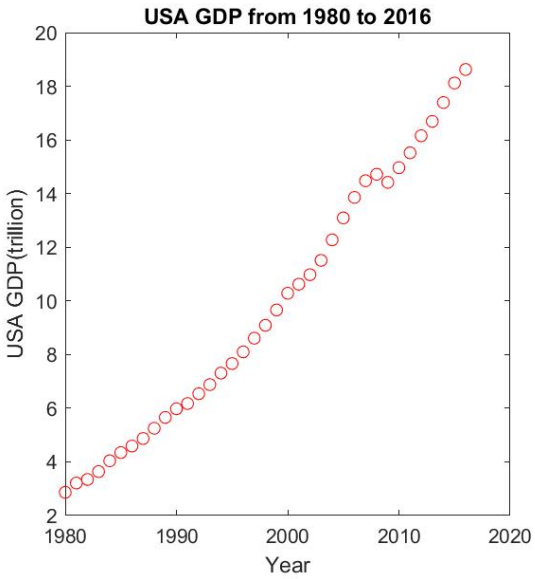
- 中国协方差矩阵系数相对于美国来说相关性较弱, 但二者均**接近于1**, 证明数据具有一定的**相关性**

2. 绘制初始图像

```
t = figure(1);
t.Position=[400,200,900,400];
subplot(1,2,1);
plot(y,CH,'b*');
```

```
5
6  title('China GDP from 1980 to 2016');
7  xlabel('Year');
8  ylabel('CHINA GDP(trillion)');
9
10 subplot(1,2,2);
11 plot(y,US,'ro');
12
13 title('USA GDP from 1980 to 2016');
14 xlabel('Year');
15 ylabel('USA GDP(trillion)');
16
```





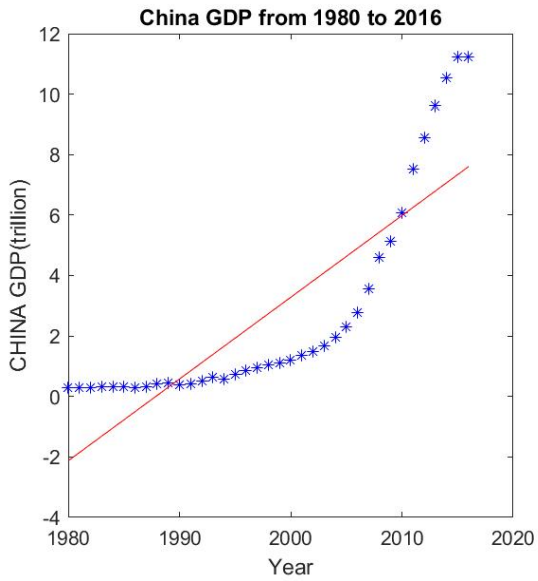
3. 多项式拟合及残差分析

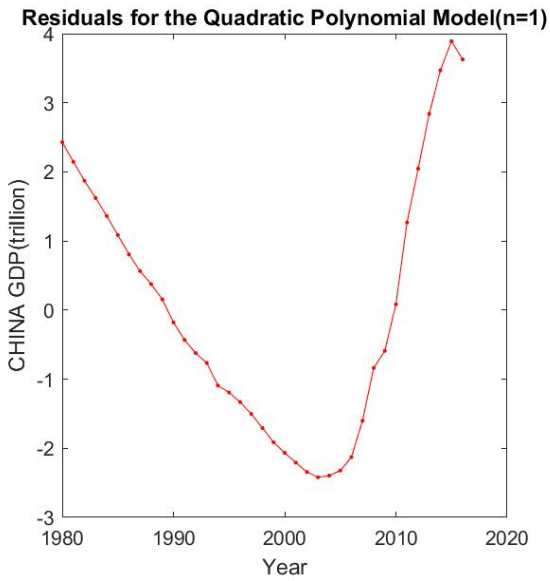
- 这里只贴出中国的代码（美国与此类似）

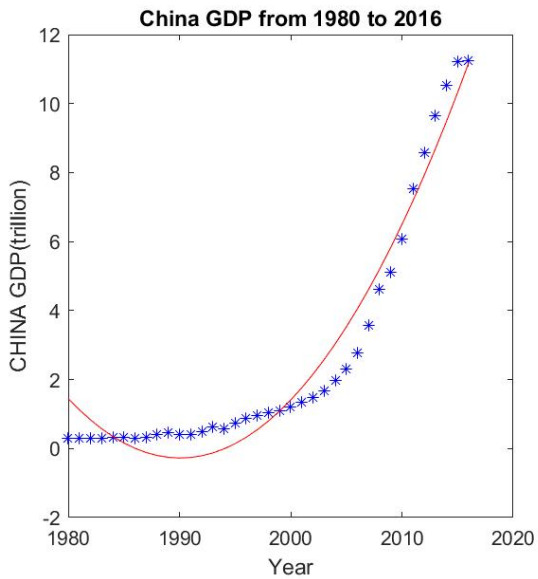
```
1  for i = 1:4
2
3      [p,~,~]= polyfit(y,CH,i);
4
5      CH_f = polyval(p,y,[],CH_mu);
6      pk = pk + 1;
7      f_pk = figure(pk);
8      f_pk.Position = [400,200,900,400];
9      subplot(1,2,1);
10
11     plot(y,CH,'b*',y,CH_f,'r-');
12     title('China GDP from 1980 to 2016');
13     xlabel('Year');
14     ylabel('CHINA GDP(trillion)');
15
16     subplot(1,2,2);
17     res = CH - CH_f;
18
```

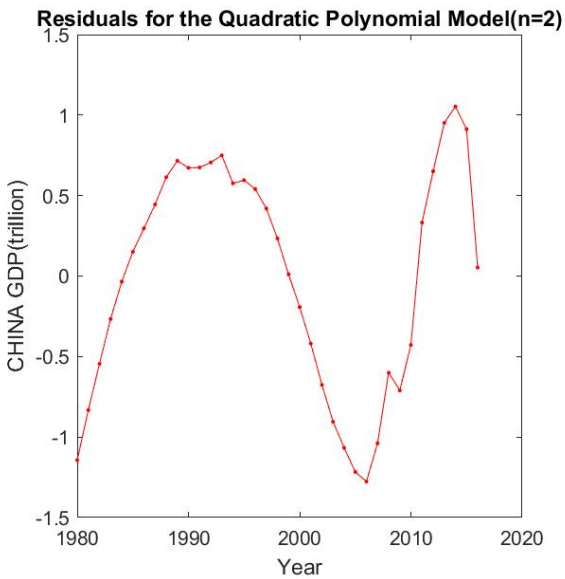
```
19
20
21      plot(y,res,'r.-');
22      title(['Residuals for the Quadratic Polynomial Model','(n=',num2str(i),')']);
23      xlabel('Year');
24      ylabel('CHINA GDP(trillion)');
25
26      CH_SSE(i) = sum(res.^2);
27      CH_SST(i) = (length(CH)-1)*var(CH);
28      CH_R_sqr(i) = 1-CH_SSE(i)/CH_SST(i);
29
30      str = strcat('China_poly(n=',num2str(i),').jpg');
31      saveas(gcf,str);
32
33  end
```

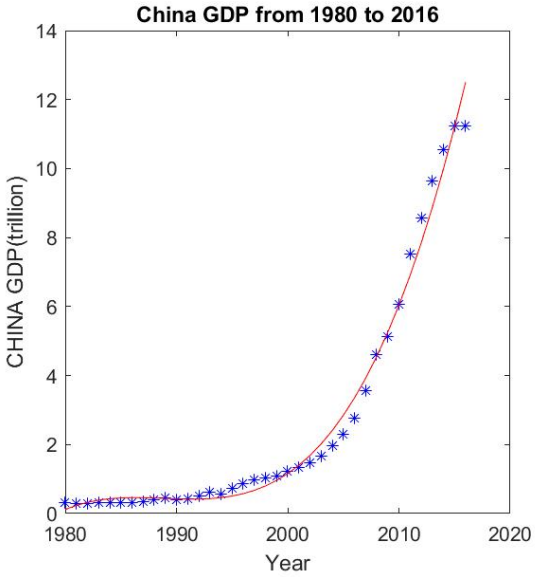
- **China**

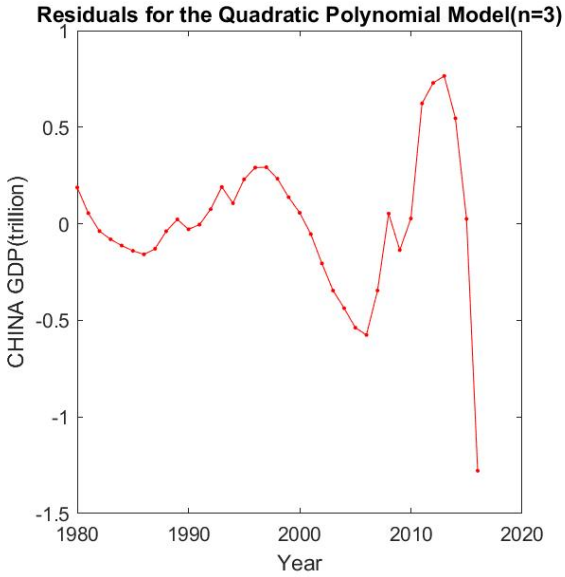


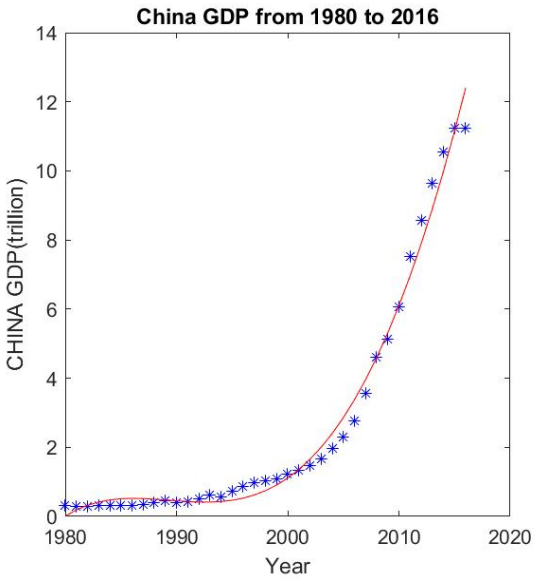


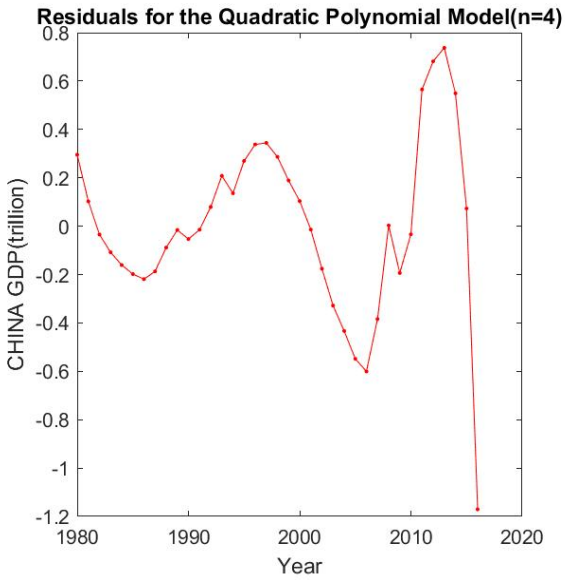


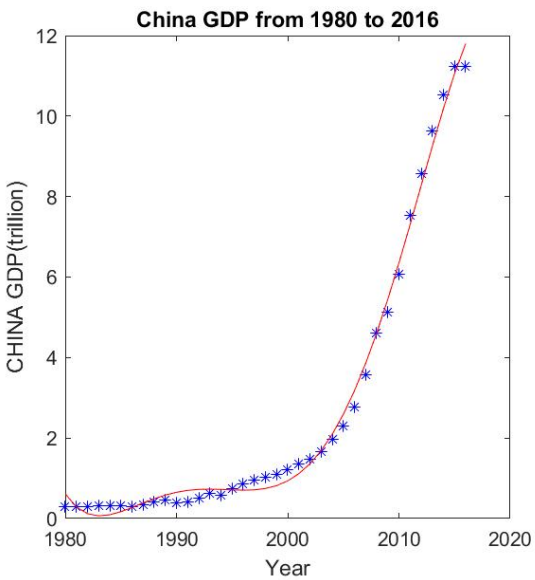


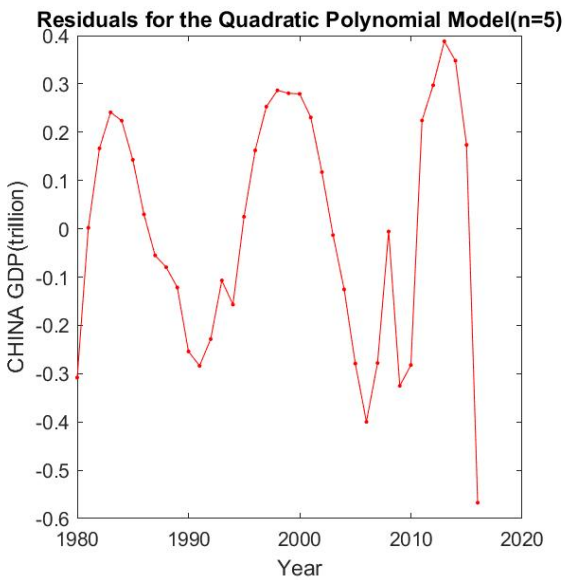

China GDP from 1980 to 2016
CHINA GDP(trillion)
Year


Residuals for the Quadratic Polynomial Model(n=3)
CHINA GDP(trillion)
Year

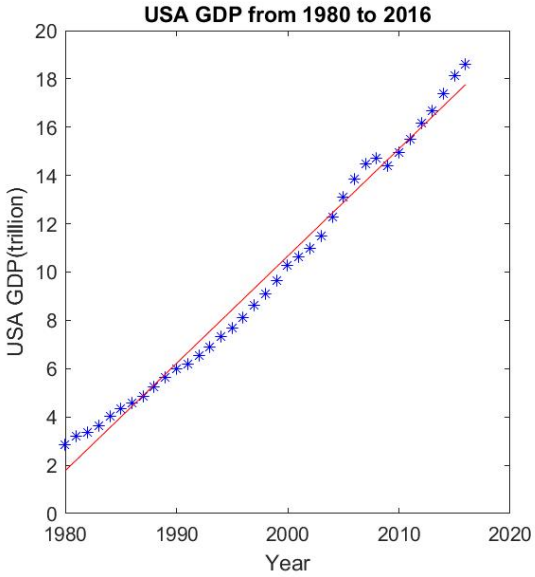

China GDP from 1980 to 2016
CHINA GDP(trillion)
Year

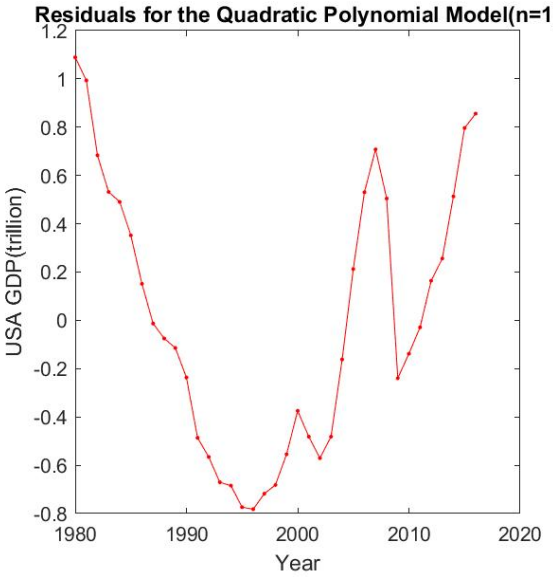

Residuals for the Quadratic Polynomial Model(n=4)
CHINA GDP(trillion)
Year

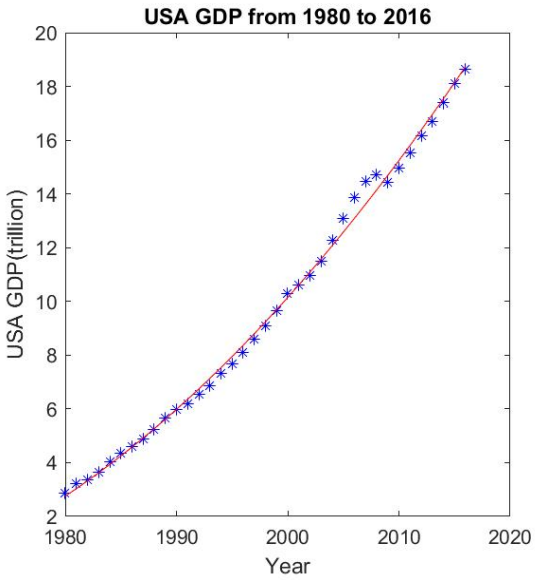

China GDP from 1980 to 2016
CHINA GDP(trillion)
Year

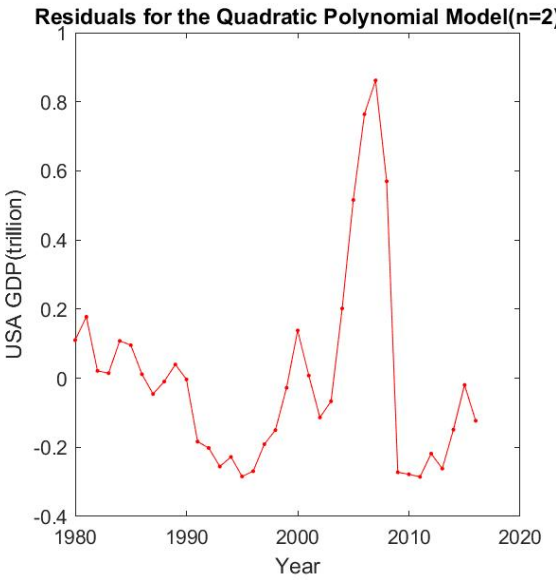

Residuals for the Quadratic Polynomial Model(n=5)
CHINA GDP(trillion)
Year

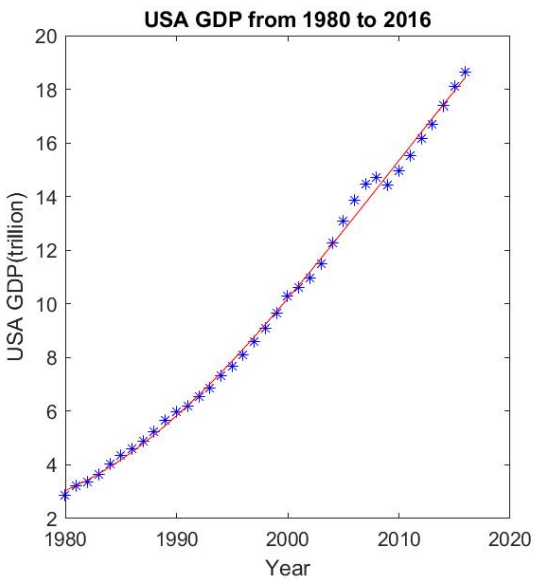
- **USA**

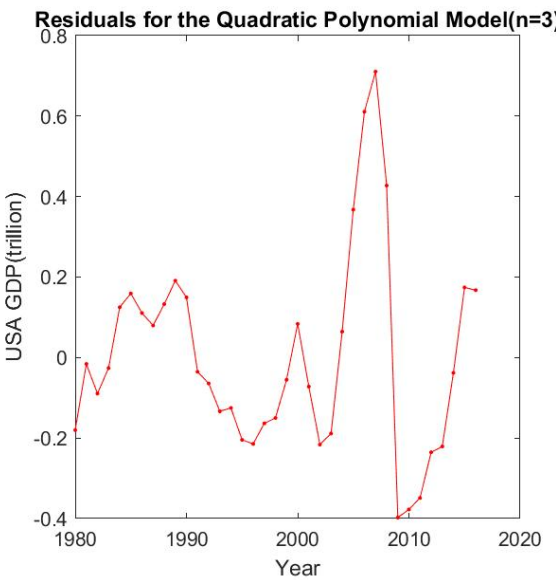

USA GDP from 1980 to 2016
USA GDP(trillion)
Year

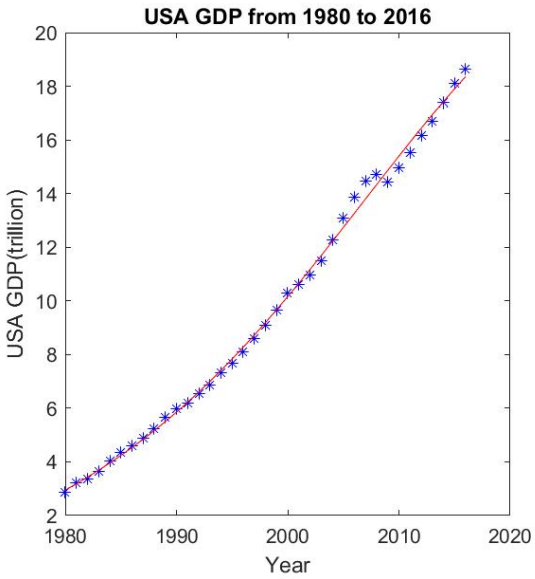

Residuals for the Quadratic Polynomial Model(n=1)
USA GDP(trillion)
Year

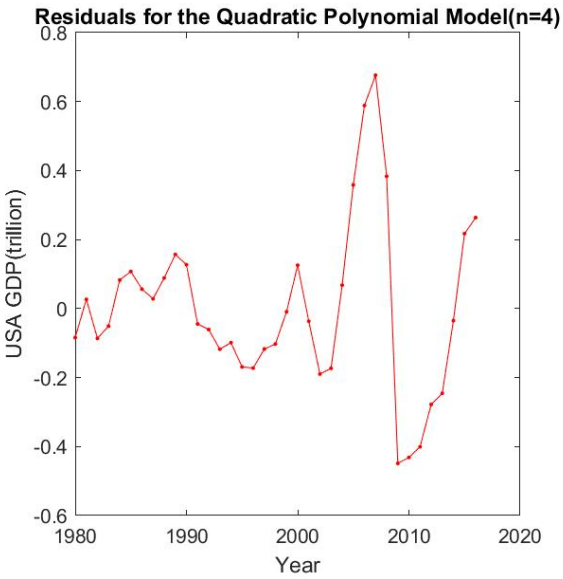

USA GDP from 1980 to 2016
USA GDP(trillion)
Year

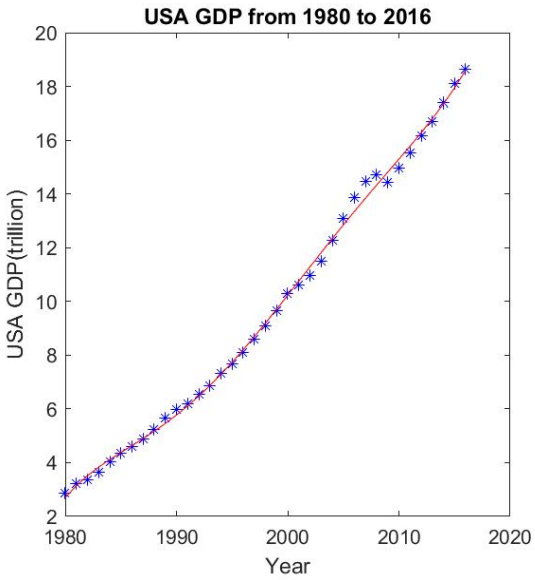

Residuals for the Quadratic Polynomial Model(n=2)
USA GDP(trillion)
Year

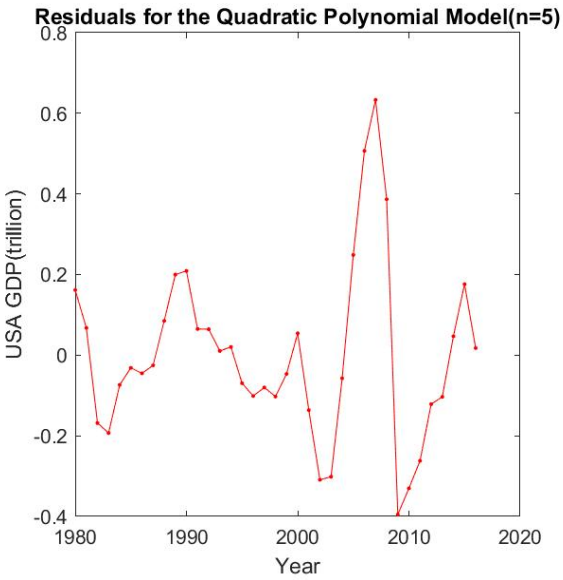

USA GDP from 1980 to 2016
USA GDP(trillion)
Year


Residuals for the Quadratic Polynomial Model(n=3)
USA GDP(trillion)
Year









- **分析过程**
  - **China**

    计算相关拟合数据

Fit name	SSE	R-squ	RMSE
China_fit(n=1)	129.207	0.705	1.921
China_fit(n=2)	17.993	0.959	0.727
China_fit(n=3)	5.033	0.988	0.391
China_fit(n=4)	4.951	0.989	0.393
China_fit(n=5)	2.185	0.995	0.266

当多项式阶数越高，拟合的准确度是在不断提高的，但是这并不代表该数学模型具有很好的预测和泛化能力。例如 n=5 和 4 时，拟合曲线明显出现了过拟合现象， 舍去。

综上，我们确定了 n = 3 时的拟合曲线。

即 $f(x) = 0.7881x^3 + 1.994x^2 + 1.547x + 0.7991$

- **USA**

  计算相关拟合数据

Fit name	SSE	R-squ	RMSE
USA_ploy(n=1)	11.198	0.987	0.566
USA_ploy(n=2)	2.854	0.997	0.290
USA_ploy(n=3)	2.237	0.997	0.260
USA_ploy(n=4)	2.172	0.997	0.261
USA_ploy(n=5)	1.713	0.998	0.235

如上，可以看到拟合数据，拟合的准确度是随着阶数的增高在不断提高的。根据残差图及拟合指标分析

我们最终选取了 n = 3 时的拟合曲线。

即

$$f(x) = -0.172x^3 + 0.5463x^2 + 5.11x + 9.239$$

4. 最小二乘法拟合

```
y = [1980:2016].';

[~,~,CH_mu] = polyfit(y,CH,0);
[~,~,US_mu] = polyfit(y,US,0);
f = @(p,y) p(1)*y.^3 + p(2)*y.^2 + p(3)*y + p(4);


y = (y-CH_mu(1))/CH_mu(2);

CH_p = lsqcurvefit(f,[1,0,0,0],y,CH);

USA_p = lsqcurvefit(f,[1,0,0,0],y,US);

```

- 计算出的结果与poly 拟合的结果**相差不大**

5. 预测2017年至2020年中、美GDP

- **中国GDP预测**

年份	预测值	实际值
2017	123,104	13.9217
2018	138,948	15.4375
2019	142,799	17.0619
2020	147,227	18.7987

- **美国GDP预测**

年份	预测值	实际值
2017	195,430	18.9624
2018	206,119	19.4616
2019	214,332	19.9538
2020	209,366	20.4383

三、结论

最终得到的三阶多项式拟合的中美两国的GDP增长数学模型、均方误差MSE及拟合优度R^2为

- **中国**

  $$f(x) = 0.7881x^3 + 1.994x^2 + 1.547x + 0.7991$$

  $R^2 = 0.988$ $RMSE = 0.391$

- **美国**

  $$f(x) = -0.172x^3 + 0.5463x^2 + 5.11x + 9.239$$

  $R^2 = 0.997$ $RMSE = 0.260$