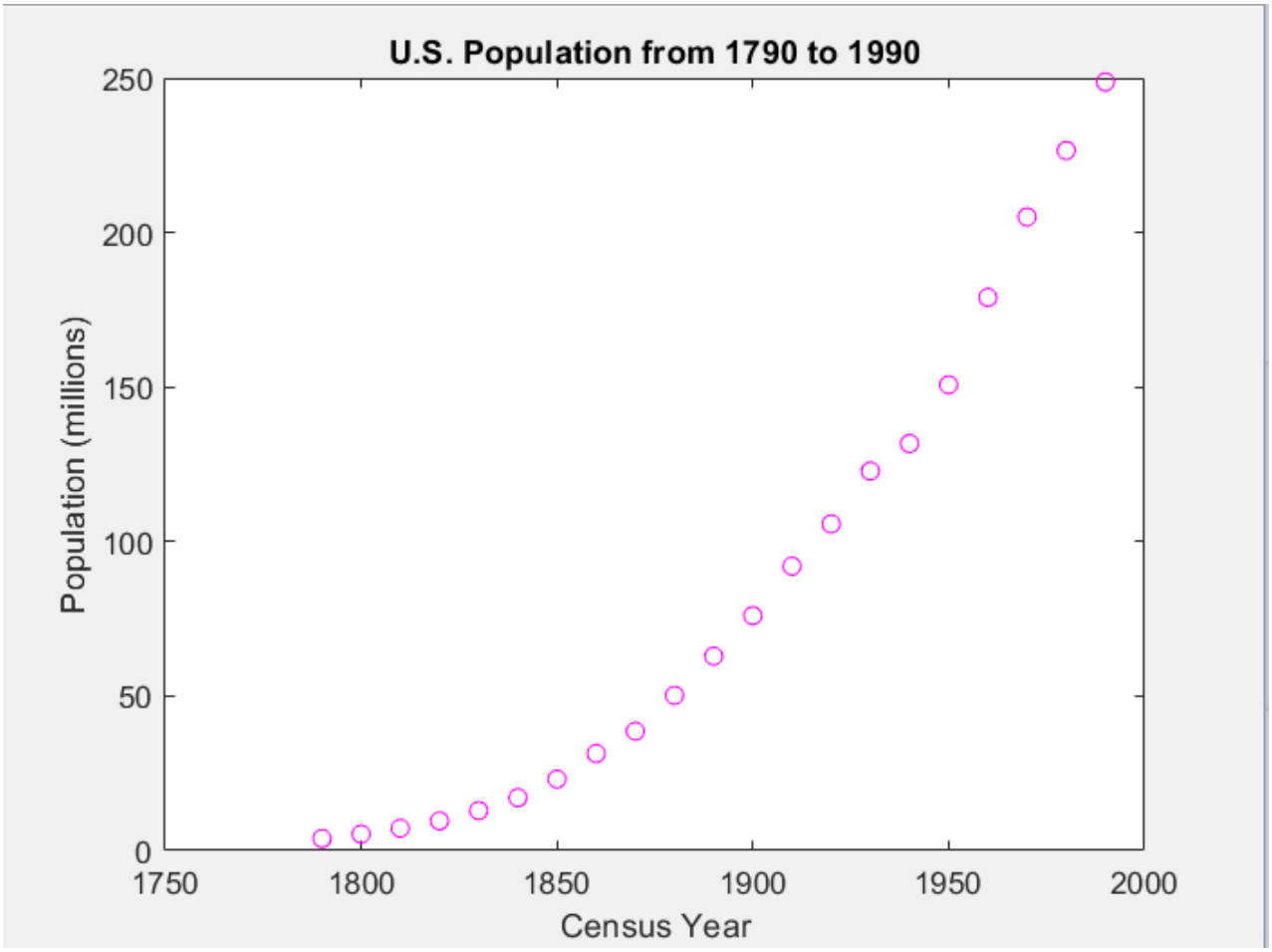
多项式曲线拟合

1. 相关性分析

- **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
- **分析过程及部分代码**
  - **散点图绘制**

```
plot(cdate,pop,'om')
title('U.S. Population from 1790 to 1990');
xlabel('Census Year');
ylabel('Population (millions)');
```



  - **计算协方差矩阵**

```
>> corrcoef(cdate,pop)

    ans =

      1.0000    0.9597
      0.9597    1.0000
```

- **结论**

1. 从图像上可以直观地看出原始样本数据具有较强的相关性
2. 协方差矩阵系数都**接近于1**,进一步证明数据的强相关性

2. 多项式拟合

- **原理**
  - [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。

      $$X^* = \frac{X - \bar{X}}{\sigma}$$

    - ```
      (cdate-mean(cdate))./std(cdate);
      //代码演示
      ```

    - 工具箱内勾选"**Center and Scale**"
  - **通过Curve Fitting Tool工具箱绘制拟合图像和残差图**
- **部分代码和图像**
  - **初始化数据**
  - ```
    %% Initialization.

    % Initialize arrays to store fits and goodness-of-fit.
    fitresult = cell( 2, 1 );
    gof = struct( 'sse', cell( 2, 1 ), ...
      'rsquare', [], 'dfe', [], 'adjrsquare', [], 'rmse', [] );

    ```
  - **拟合曲线(注:这里只贴出poly2的源码)**

    ```
    %% Fit: 'poly2'.
    [xData, yData] = prepareCurveData( cdate, pop );

    ```

```
% Set up fittype and options.
ft = fittype( 'poly2' );

% Fit model to data.
[fitresult{1}, gof(1)] = fit( xData, yData, ft, 'Normalize', 'on' );

% Create a figure for the plots.
figure( 'Name', 'poly2' );

% Plot fit with data.
subplot( 2, 1, 1 );
h = plot( fitresult{1}, xData, yData );
legend( h, 'pop vs. cdate', 'poly2', 'Location', 'NorthEast' );
% Label axes
xlabel cdate
ylabel pop
grid on

% Plot residuals.
subplot( 2, 1, 2 );
h = plot( fitresult{1}, xData, yData, 'residuals' );
legend( h, 'poly2 - residuals', 'Zero Line', 'Location', 'NorthEast' );

% Label axes
xlabel cdate
ylabel pop
grid on
```

- **poly2~poly6,exp2的拟合曲线及残差图**

3. 拟合优度分析

原理

1. **SSE(残差平方和): the Sum of Square due to Error**

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

   该参数越接近于0,表示曲线拟合越成功

2. **RMSE(均方根误差、标准差)：Root mean squared error**

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

   每个误差对 RMSE 的影响与平方误差的大小成正比；因此较大的误差对 RMSE 的影响非常大,所以RMSE 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

> $MSE = \frac{SSE}{n}$

3. **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 // **区分于相关系数r和**$\rho\ xy$

   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.

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

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

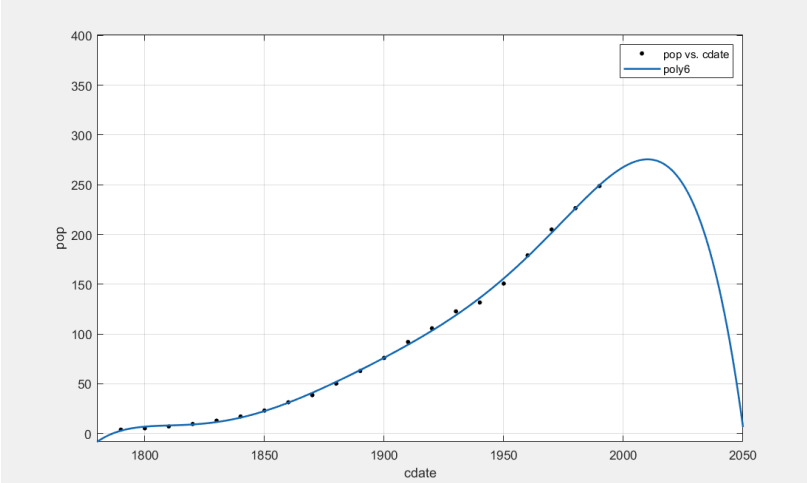
> $SS\ reg = \sum_{i=1}^{n} (\hat{y} - \bar{y})^2$
>
> $SS\ res = \sum_{i=1}^{n} (y_i - \hat{y})^2 = e_i^2$
>
> $SS\ tot = \sum_{i=1}^{n} (y_i - \bar{y})^2$
>
> 回归平方和：**SSR(Sum of Squares for regression) = ESS (explained sum of squares)**
>
> 残差平方和：**SSE(Sum of Squares for Error) = RSS(residual sum of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

4. $R^2\ adj$ **(校正决定系数): adjust R-square**

   R^2 评价拟合模型的好坏具有一定的局限性，即使向模型中增加的变量没有统计学意义，R^2值仍会增大。因此需对其进行校正，从而形成了校正的决定系数(Adj R^2) 。与R^2 不同的是，当模型中增加的变量没有统计学意义时，Adj R^2会减小，因此Adj R^2 是衡量所建模型好坏的重要指标之一，**Adj R^2 越大，模型拟合得越好。**

> 校正决定系数（Adj R^2）引入了**样本数量**和**特征数量**，公式如下：
>
> $R^2\ adj = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$

- # 分析过程及部分代码
  - **拟合人口普查数据的目的是预测出未来人口数**
    - 对于poly6来说,数据明显超出了我们想要得到的结果,因此ploy6舍去

- 

  - **注: 其余拟合图像均有着合理的预测结果,在此不再赘述.**
- **过拟合判断**

> 得到拟合结果后, 若计算出来的最高项系数过0(Zero Crossing)并且在0附近,则表明这个系数对于真实的**多项式拟合**没有任何的帮助, 即发生了**过拟合**(overfitting)

  - 观察到poly5的数据**p1,p2,p3与0非常接近的且p3<0,**证明ploy5拟合曲线发生了过拟合,故ploy5舍去
  - Results

Linear model Poly5:
f(x) = p1*x^5 + p2*x^4 + p3*x^3 + p4*x^2 + p5*x + p6
where x is normalized by mean 1890 and std 62.05
Coefficients (with 95% confidence bounds):
p1 = 0.5877 (-2.305, 3.48)
p2 = 0.7047 (-1.684, 3.094)
p3 = -0.9193 (-10.19, 8.356)
p4 = 23.47 (17.42, 29.52)
p5 = 74.97 (68.37, 81.57)
p6 = 62.23 (59.51, 64.95)

Goodness of fit:
SSE: 144.2
R-square: 0.9988
Adjusted R-square: 0.9984
RMSE: 3.1

  - **注: 其余拟合参数均无上述情况,在此不再赘述.**

- **工具箱计算部分数据如下表:**

Fit Name	SSE	RMSE	R-square	DFE	Adj R-sq
poly6	106.9276	2.764	0.99913	14	0.998764
poly5	144.1661	3.100	0.99883	15	0.998444
poly4	145.9689	3.020	0.99882	16	0.998523
poly3	149.7687	2.968	0.99879	17	0.998574
poly2	159.0293	2.972	0.99871	18	0.99857
exp2	475.9491	5.291	0.99615	17	0.995468

- 由表可得

1. SSE最大的为exp2,拟合效果**极差**,exp2舍去
2. 对比剩余曲线,其R^2 和 Adj R^2 近似相等,故对比RMSE

Fit Name	SSE	RMSE	R-square	DFE	Adj R-sq
poly3	149.7687	2.968	0.99879	17	0.998574
poly2	159.0293	2.972	0.99871	18	0.99857
poly4	145.9689	3.020	0.99882	16	0.998523

- 显然poly3>poly2>poly4,ploy2和ploy4舍去

4.数学模型

由以上分析，拟合次数为3时多项式系数分别为 0.921，25.183，73.859，61.744可得到最终的三阶多项式拟合的美国人口增长数学模型

$$f(x) = 0.921x^3 + 25.183x^2 + 73.859x + 61.744$$