Short introduction, what is the data what is the task.

# 1 Aligning the dates for the three datasets

```
head:
    year  num_mortage  num_possession  Unemployment rate  IR rate
0  1975    5076000.0          4870.0                4.5  11.0000
1  1976    5322000.0          4950.0                5.4  11.1137
2  1977    5582000.0          4680.0                5.6   8.8772
tail:
     year  num_mortage  num_possession  Unemployment rate  IR rate
36  2011   11384000.0         37300.0                8.1      0.5
37  2012   11284000.0         33900.0                8.0      0.5
38  2013   11186000.0         28900.0                7.6      0.5
```

Figure 1: head and tails of the aligned dataset

maybe visualize the data.

| | num_moratage | num_possession | Unemployment rate |
|---|---|---|---|
| IR_rate | -0.760988 | -0.285785 | 0.242834 |

repossession rate num-poss num-mort

Table 1: coefficients of correlation of IR_rate with independent variables

From table 1 we observed a strong negative correlation between IR_rate and "num_mortage". In following sections, the independent variables are "num_mortage", "num_ possession", "Unemployment rate", the dependent variable is "IR rate", we hope to model the relationship between independent and dependent variables using simple regression models, polynomial regression models and piecewise polynomial regression models.

# 2 Simple Regression Models

| Model name | $R^2$ | Log-likelihood | AIC | BIC |
|---|---|---|---|---|
| SR | 0.598 | -93.853 | 195.7 | 202.4 |
| SR_updated | 0.598 | -93.853 | 193.7 | 198.7 |

Table 2: Model quality quantities table for SR and SR_updated

what if you do the regression only with this variable

For SR_updated model, we removed the "Unemployment rate" columns from the predictors as SR model summary suggests $P > |t|$ of this coefficient is 0.988 hence this column is highly likely irrelevant. From the AIC and BIC value in table 1, the SR_updated is better.

# 3 Polynomial Regression Models

We observed from table 3 that the best model is adding every predictor raised to power of 2, and orthogonalising them.

↳ specify the model more detailed

# 4 Piecewise Polynomial Regression Models

We picked 4 knots for each predictor for the piecewise polynomial regression by using quantile method. From Table 4 and Figure 4 we can see the this PPR outperforms all previous models in every quantity chosen, though the BIC of PPR is still lower than the of PR_2_ortho, we slightly worry about the problem of overfitting.

↳ Good, some more comparison and conclusion would be good as well as interpretations.

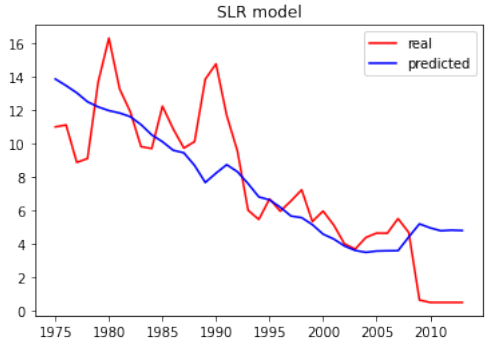
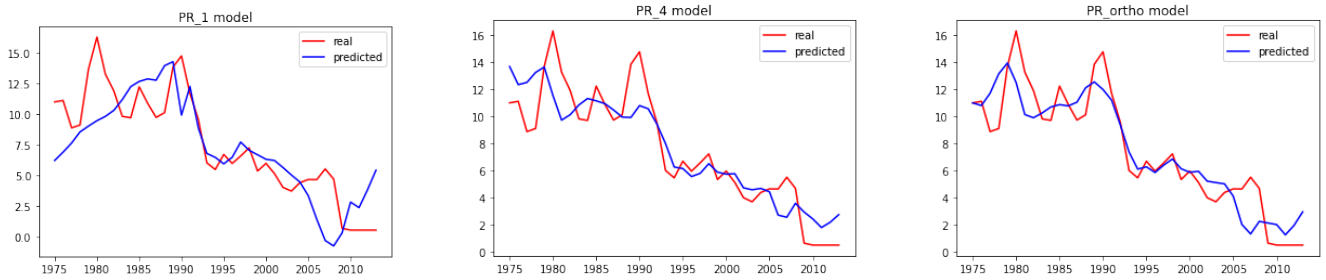



Figure 2: Plotting of the real and predicted IR_rate using SR_updated model against year

| Model name | raised power | orthogonalised? | $R^2$ | Log-likelihood | AIC | BIC |
|---|---|---|---|---|---|---|
| PR_4 | 4 | No | 0.544 | -96.306 | 198.6 | 203.6 |
| PR_3 | 3 | No | 0.726 | -86.386 | 180.8 | 187.4 |
| PR_2 | 2 | No | 0.777 | -82.351 | 176.7 | 186.7 |
| PR_2_ortho | 2 | Yes | 0.824 | -77.733 | 169.5 | 181.1 |

Table 3: Model quality quantities table for various polynomial regression models




(a) PR_4 (b) PR_2 (c) PR_2_orthogonal

Figure 3: plottings with different polynomial regression models

| Model name | $R^2$ | Log-likelihood | AIC | BIC |
|---|---|---|---|---|
| PPR | 0.926 | -60.774 | 153.5 | 180.2 |
| SR_updated | 0.598 | -93.853 | 193.7 | 198.7 |
| PR_2_ortho | 0.824 | -77.733 | 169.5 | 181.1 |

Table 4: Model quality quantities table for PPR and best SR and PR models

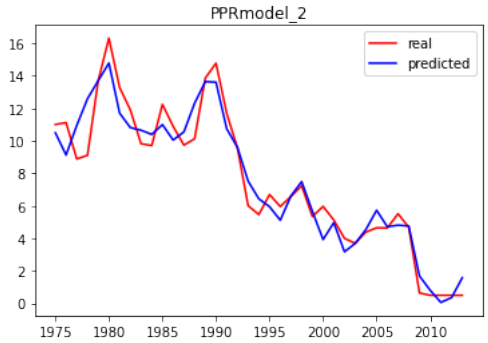



Figure 4: Plotting of the real and predicted IR_rate using PR model against year

# Q2-Aisha Xu

October 26, 2020

[21]:
```
import numpy as np
import statsmodels.api as sm
from math import *
```

[16]:
```
def knots_finder(x,deg):



    m = deg-2
    if (deg<3):
        stop("degree of freedom has to be no less than 3")
    q = np.arange(0,d,1)
    q = q/(d-1)
    q = q[1:-1]

    k = np.quantile(x,q,axis = 0)
    return k
```

[17]:
```
def pielinear(x,k,first = True):
    import numpy as np

    n = len(x)
    m = len(k)
    M = np.zeros((n,m+2))
    M[:,0] = np.ones(n)
    M[:,1] = x

    X = np.array([x for i in range(m)])
    X = X.T
    K = np.array([k for j in range(n)])
    M[:,2:] = np.maximum(X-K,np.zeros((n,m)))
    if first: #whether the keep the first columns of 1
        return M
    else:
        return M[:,1:]
```

[ ]:
```
def pielinear_multivari(x_vec,deg):
    k = knots_finder(x_vec,deg)
```

```
    dim = x_vec.shape[1]
    M = [pielinear(x_vec[:,i],k[:,i],first = False) for i in range(dim)]
    X_vec = np.concatenate(M, axis = 1)
    return X_vec
```

[20]:
```
def best_knots(y,x_vec):
    max_d = int(sqrt(x_vec.shape[0]))
    for d in range(max_d):
        X_vec = pielinear_multivari(x_vec,d+1+2)
        X_vec_new = sm.add_constant(X_vec)
        model = sm.OLS(y,X_vec_new).fit()
        aic_ = model.aic
        print(d,aic_)
        if d==0:
            aic = aic_
        if d != 0 and aic_ < aic:
            aic = aic_
            m_op = d+1
    return m_op
#m_op is the optimal number of knots using criteria AIC
```

Assume: $\eta_i, \varepsilon_i$ are independent standard normal r.v's

$\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n, \eta_1, \eta_2, \cdots, \eta_n$ ~~are~~ i.i.d $N(0,1)$.

(Hope this is what the question means)

Let $z = \beta_1\sigma_1\eta + \sigma_2\varepsilon$

As $\eta, \varepsilon \overset{i.i.d.}{\sim} N(0,1)$.

$z \sim N(0, (\beta_1\sigma_1)^2 + \sigma_2^2)$. , $z = Y - \beta_0 - \beta_1 X$

$$l = \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi[(\beta_1\sigma_1)^2+\sigma_2^2]}} e^{-\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2[(\beta_1\sigma_1)^2+\sigma_2^2]}} \text{ 未.}$$

$$= \sum_{i=1}^{n} -\frac{1}{2}\log 2\pi\left[(\beta_1\sigma_1)^2+\sigma_2^2\right] - \frac{(y_i-\beta_0-\beta_1 x_i)^2}{2[(\beta_1\sigma_1)^2+\sigma_2^2]}$$

$\frac{\partial l}{\partial \beta_0}$

$$= -\frac{n}{2}\log\left(2\pi\left[(\beta_1\sigma_1)^2+\sigma_2^2\right]\right) - \frac{\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2}{2[(\beta_1\sigma_1)^2+\sigma_2^2]}$$

$$= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\left[(\beta_1\sigma_1)^2+\sigma_2^2\right] - \frac{\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2}{2[(\beta_1\sigma_1)^2+\sigma_2^2]}$$

$$\frac{\partial l}{\partial \beta_0} = \frac{\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)}{2[(\beta_1\sigma_1)^2+\sigma_2^2]} = \frac{\Sigma y_i - n\beta_0 - \beta_1\Sigma x_i}{2[(\beta_1\sigma_1)^2+\sigma_2^2]}$$

$$\frac{\partial l}{\partial \beta_1} = -\frac{n}{2}\frac{2(\beta_1\sigma_1)\sigma_1}{(\beta_1\sigma_1)^2+\sigma_2^2} + \frac{\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)x_i}{[(\beta_1\sigma_1)^2+\sigma_2^2]}$$

$$+ \frac{\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2}{2[(\beta_1\sigma_1)^2+\sigma_2^2]^2} \cdot 2(\beta_1\sigma_1)\sigma_1$$

$\frac{\partial l}{\partial \beta_1} = 0 \Rightarrow$

$$-n\beta_1\sigma_1^2\left[(\beta_1\sigma_1)^2+\sigma_2^2\right] + \left(\sum_{i=1}^{n} y_i x_i - \beta_0\Sigma x_i - \beta_1\Sigma x_i^2\right)\left[(\beta_1\sigma_1)^2+\sigma_2^2\right]$$

$$+ \beta_1\sigma_1^2\left(\Sigma y_i^2 + n\beta_0^2 + \beta_1^2\Sigma x_i^2 - 2\beta_0\Sigma y_i - 2\beta_1\Sigma x_i y_i + 2\beta_0\beta_1\Sigma x_i\right) = 0$$

$\frac{\partial l}{\partial \beta_0} = 0 \Rightarrow \quad \Sigma y_i - n\beta_0 - \beta_1\Sigma x_i = 0$

$$\beta_0 = \frac{\Sigma y_i - \beta_1\Sigma x_i}{n}$$

Q3

$$\underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \underbrace{\begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1m} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nm} \end{pmatrix}}_{:= X} \underbrace{\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix}}_{:= \underline{\beta}} + \sigma \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{:= \underline{\varepsilon}}$$

$\underline{\varepsilon} \sim N(\underline{0}, \Sigma)$. where $\Sigma_{ij} = \rho^{|i-j|}$ (since $\mathrm{Var}(\varepsilon_i) = 1$ for any $i$).

$$Y \sim N(X\underline{\beta}, \sigma^2 \Sigma)$$

$$l = \log f(y_1, \cdots y_n) = \log \frac{1}{(2\pi)^{\frac{n}{2}}} \frac{1}{\sqrt{\sigma^2 \det \Sigma}} e^{-\frac{(\underline{y} - X\underline{\beta})^T \Sigma^{-1} (\underline{y} - X\underline{\beta})}{2\sigma^2}}$$

$$= \log \frac{1}{(2\pi)^{\frac{n}{2}} \sqrt{\sigma^2 \det \Sigma}} - \frac{(\underline{y} - X\underline{\beta})^T \Sigma^{-1} (\underline{y} - X\underline{\beta})}{2\sigma^2}$$

$$\frac{\partial l}{\partial \underline{\beta}} = \frac{\partial}{\partial \underline{\beta}} \left( \frac{\underline{y}^T \Sigma^{-1} \underline{y} - 2\underline{\beta}^T X^T \Sigma^{-1} \underline{y} + \underline{\beta}^T X^T \Sigma^{-1} X \underline{\beta}}{2\sigma^2} \right)$$

$$= \frac{X^T \Sigma^{-1} \underline{y}}{\sigma^2} - \frac{X^T \Sigma^{-1} X \underline{\beta}}{\sigma^2} = 0$$

$$\Rightarrow \quad \hat{\underline{\beta}} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} \underline{y}$$

$$-n^2\beta_1\sigma_1^2[\beta_1^2+\lambda] + (n\Sigma y_i x_i - \Sigma y_i \Sigma x_i + \beta_1(\Sigma x_i)^2 - n\beta_1\Sigma x_i^2)(\beta_1^2+\lambda)$$

$$+\beta_1\left(n\Sigma y_i^2 + (\Sigma y_i)^2 - 2\beta_1 \Sigma y_i \Sigma x_i + \beta_1^2(\Sigma x_i)^2 + n\beta_1^2\Sigma x_i^2 \right.$$
$$\left. - 2(\Sigma y_i)^2 + 2\beta_1 \Sigma x_i \Sigma y_i - 2n\beta_1 \Sigma x_i y_i + 2n\beta_0\beta_1 \Sigma x_i\right) = 0.$$

$$2\beta_1\Sigma x_i(\Sigma y_i - \beta_1 \Sigma x_i)$$
$$= 2\beta_1 \Sigma x_i \Sigma y_i - 2\beta_1^2(\Sigma x_i)^2.$$

$$-n^2\beta_1\sigma_1^2(\beta_1^2+\lambda) + (n\Sigma y_i x_i - \Sigma y_i \Sigma x_i)(\beta_1^2+\lambda) + \beta_1(\beta_1^2+\lambda)((\Sigma x_i)^2 - n\Sigma x_i^2)$$
$$+\beta_1\left[n\Sigma y_i^2 - (\Sigma y_i)^2\right] + \beta_1^3\left[(\Sigma x_i)^2 - n\Sigma x_i^2\right]$$
$$+\beta_1^3\left[n\Sigma x_i^2 - (\Sigma x_i)^2\right] + 2\beta_1^2(\Sigma x_i \Sigma y_i - n\Sigma x_i y_i) = 0.$$

$\Rightarrow$

$$\left[(\Sigma x_i)^2 - n\Sigma x_i^2\right](\beta_1^3 + \lambda\beta_1 - \beta_1^3) + (\Sigma x_i y_i - n\Sigma x_i y_i)(2\beta_1^2 - \beta_1^2 - \lambda)$$
$$+\beta_1\left[n\Sigma y_i^2 - (\Sigma y_i)^2\right] - n^2\beta_1\sigma_1^2(\beta_1^2+\lambda) = 0.$$

$$\left[\lambda\left[(\Sigma x_i)^2 - n\Sigma x_i^2\right] + \left(n\Sigma y_i^2 - (\Sigma y_i)^2\right)\right]\beta_1 + (\Sigma x_i y_i - n\Sigma x_i y_i)(\beta_1^2 - \lambda)$$
$$-n^2\beta_1\sigma_1^2(\beta_1^2+\lambda) = 0.$$

$$\Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - n\bar{x}^2.$$
$$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = \Sigma x_i y_i - \bar{x}\bar{y}$$

$$\beta_0 = \bar{y} - \beta_1\bar{x}$$

$$z = y - \beta_0\begin{pmatrix}1\\ \vdots\end{pmatrix} - \beta_1 x.$$

$$Z = (\Sigma y_i x_i - (\bar{y} - \beta_1\bar{x})\Sigma x_i - \beta_1 \Sigma x_i^2).$$

$$\Sigma y_i x_i + \beta_1 \bar{x}$$