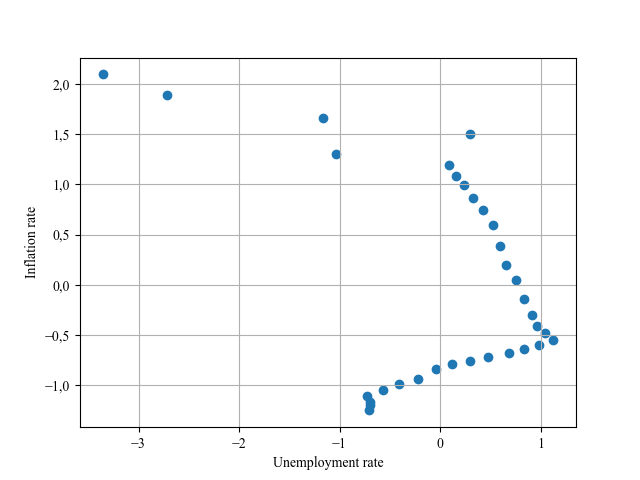
Кривая Филлипса для страны Индия

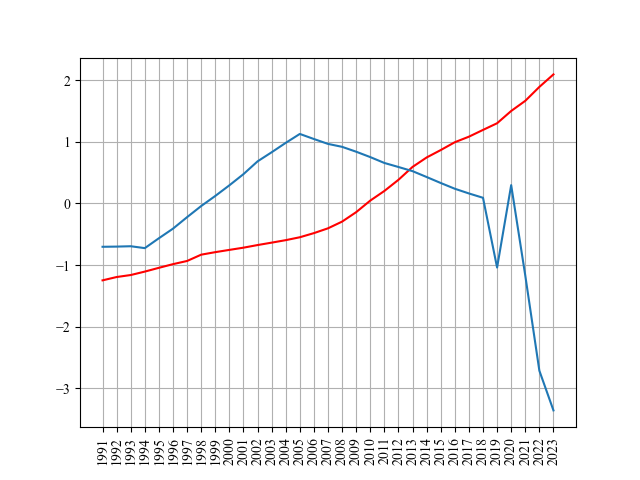
Получим данные о динамике безработицы(Inflation rate (%)) и динамике безработицы(Unemployment rate (%)) из статистики Мирового Банка о стране Индия.

|  |  |  |
| --- | --- | --- |
| date | Inflation rate | Unemployment rate |
| 1991 | 26.1320914258645 | 6.85 |
| 1992 | 29.2124945523449 | 6.853 |
| 1993 | 31.0607370914259 | 6.859 |
| 1994 | 34.2438214116532 | 6.828 |
| 1995 | 37.7452131691141 | 6.99 |
| 1996 | 41.1336584557082 | 7.147 |
| 1997 | 44.0805774514448 | 7.335 |
| 1998 | 49.9128076740881 | 7.517 |
| 1999 | 52.2436461392705 | 7.682 |
| 2000 | 54.3383216485078 | 7.856 |
| 2001 | 56.3919261013738 | 8.039 |
| 2002 | 58.815172903837 | 8.248 |
| 2003 | 61.0535954523922 | 8.397 |
| 2004 | 63.3536380862151 | 8.551 |
| 2005 | 66.0438512553292 | 8.697 |
| 2006 | 69.8720985315016 | 8.614 |
| 2007 | 74.3249644718143 | 8.534 |
| 2008 | 80.5305542396968 | 8.486 |
| 2009 | 89.2941733775462 | 8.406 |
| 2010 | 100 | 8.318 |
| 2011 | 108.911793364834 | 8.222 |
| 2012 | 119.235538897084 | 8.156 |
| 2013 | 131.18041028234 | 8.088 |
| 2014 | 139.924446113916 | 7.992 |
| 2015 | 146.790501522574 | 7.894 |
| 2016 | 154.054013105394 | 7.8 |
| 2017 | 159.18119775209 | 7.723 |
| 2018 | 165.451068899504 | 7.652 |
| 2019 | 171.621576003377 | 6.51 |
| 2020 | 182.988822584425 | 7.859 |
| 2021 | 192.378724699015 | 6.38 |
| 2022 | 205.266241146235 | 4.822 |
| 2023 | 216.862025027426 | 4.172 |

Стандартизируем полученнные данные и выведем получившеееся распределение точек



Проведем анализ изменения инфляции и безработицы в стране.



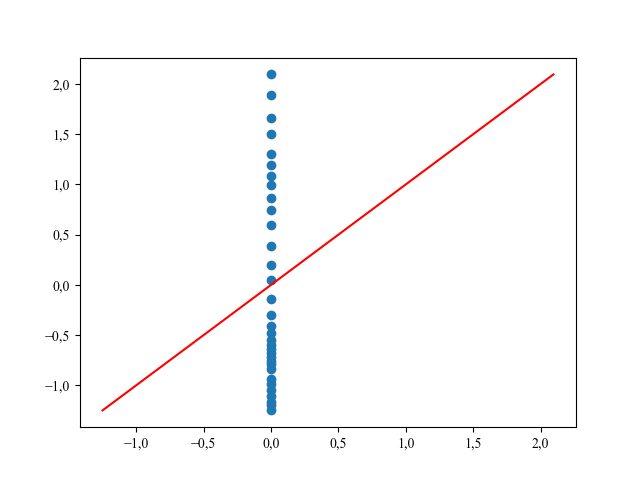
Как видно из графиков, за рассматриваемый период инфляция достигла максимального значения в 2023 году, а безработица в 2005 году, а минимумы в 1991 и 2023 годах соответственно для инфляции и безработицы.

Построим модель линейной регрессии на получившихся данных  
  
 OLS Regression Results   
==============================================================================  
Dep. Variable: y R-squared: 0.149  
Model: OLS Adj. R-squared: 0.122  
Method: Least Squares F-statistic: 5.446  
Date: Sun, 03 Nov 2024 Prob (F-statistic): 0.0263  
Time: 20:45:16 Log-Likelihood: -44.155  
No. Observations: 33 AIC: 92.31  
Df Residuals: 31 BIC: 95.30  
Df Model: 1   
Covariance Type: nonrobust   
==============================================================================  
 coef std err t P>|t| [0.025 0.975]  
------------------------------------------------------------------------------  
const -7.633e-17 0.166 -4.61e-16 1.000 -0.338 0.338  
x1 -0.3866 0.166 -2.334 0.026 -0.724 -0.049  
==============================================================================  
Omnibus: 6.909 Durbin-Watson: 0.044  
Prob(Omnibus): 0.032 Jarque-Bera (JB): 2.149  
Skew: -0.129 Prob(JB): 0.341  
Kurtosis: 1.777 Cond. No. 1.00  
==============================================================================  
  
Notes:  
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.Как видно, качество модели мало

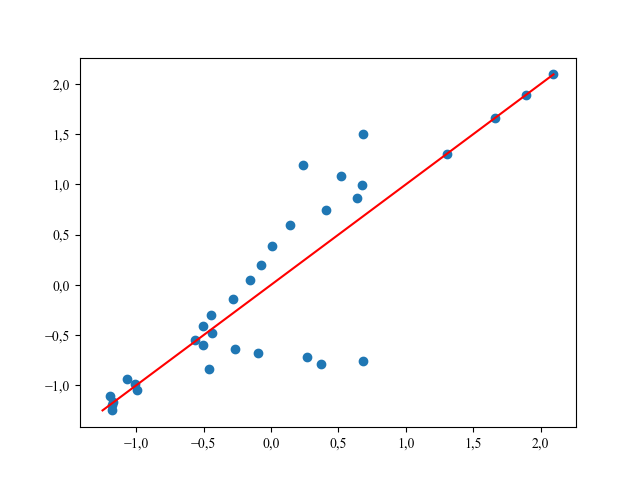
Построим модель гиперболической функции:  
  
 OLS Regression Results   
==============================================================================  
Dep. Variable: y R-squared: 0.090  
Model: OLS Adj. R-squared: 0.061  
Method: Least Squares F-statistic: 3.063  
Date: Sun, 03 Nov 2024 Prob (F-statistic): 0.0900  
Time: 20:45:16 Log-Likelihood: -45.270  
No. Observations: 33 AIC: 94.54  
Df Residuals: 31 BIC: 97.53  
Df Model: 1   
Covariance Type: nonrobust   
==============================================================================  
 coef std err t P>|t| [0.025 0.975]  
------------------------------------------------------------------------------  
x1 0.0578 0.033 1.750 0.090 -0.010 0.125  
const -0.0329 0.172 -0.191 0.850 -0.384 0.319  
==============================================================================  
Omnibus: 3.368 Durbin-Watson: 0.198  
Prob(Omnibus): 0.186 Jarque-Bera (JB): 2.969  
Skew: 0.655 Prob(JB): 0.227  
Kurtosis: 2.336 Cond. No. 5.25  
==============================================================================  
  
Notes:  
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.Как видно, качество модели также мало

Попробуем совершить полиномиальные преобразования

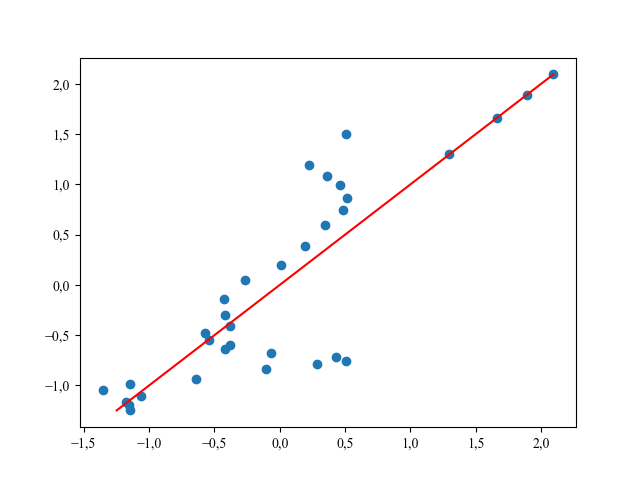
Степень полинома = 0,  
R2 значение модели = 0.0



Степень полинома = 13,  
R2 значение модели = 0.7659892546195457

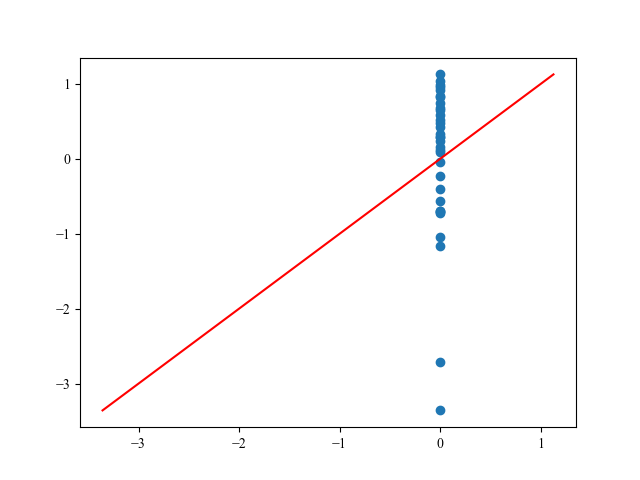


Степень полинома = 26,  
R2 значение модели = 0.7415783081424667

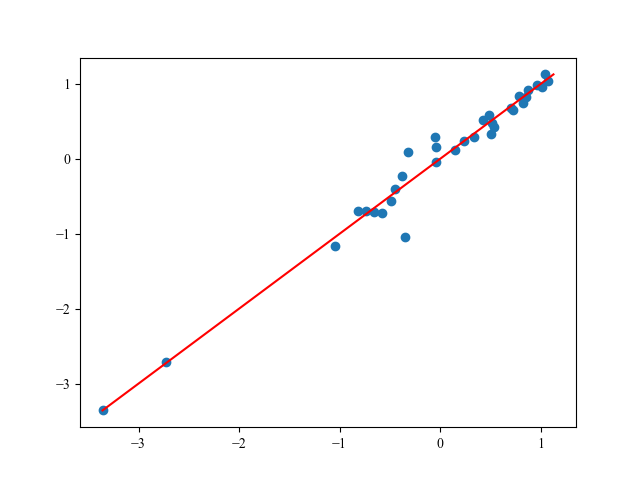


Все равно R2 счет мал.  
  
 Если смотреть на график распределения под углом в 90 градусов, то можно предположить, что это полином степени N.  
 Попробуем сменить зависимую свободную переменные местами и построить модель.

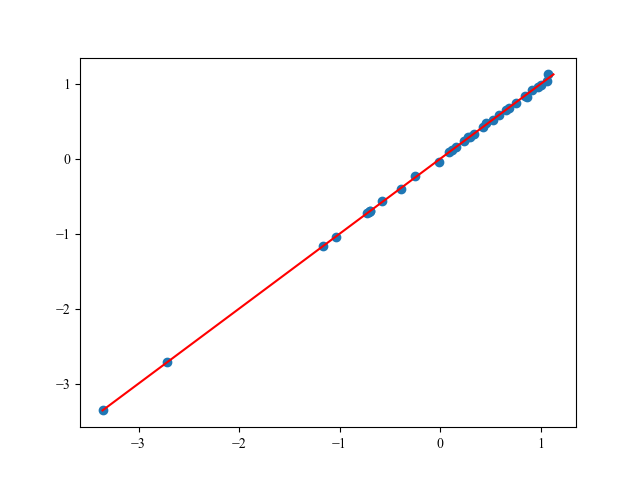
Степень полинома = 0,  
R2 значение модели = 0.0



Степень полинома = 13,  
R2 значение модели = 0.969899049060189

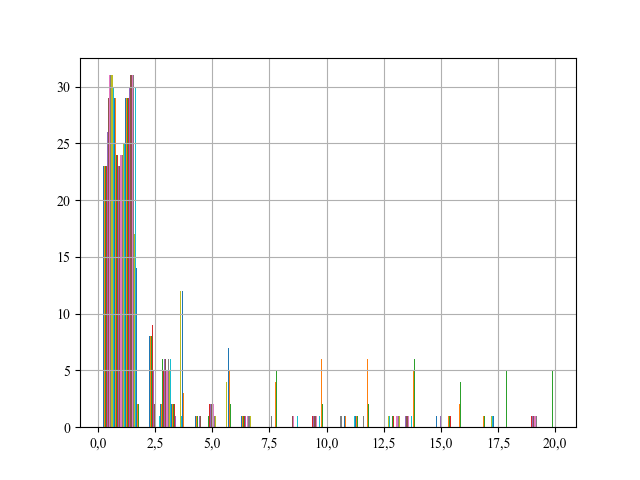


Степень полинома = 26,  
R2 значение модели = 0.9997805164378454



Мы нашли модель с R2 = 0.9997805164378454

Проведем тестирование модели.   
  
 Проверим остатки на нормальность визуально



K-S тест

Тест Колмогорова-Смирнова (или K-S тест) — это непараметрический статистический тест, применяемый для проверки соответствия распределения выборки заданному теоретическому распределению. Тест позволяет оценить, насколько эмпирическое распределение данных совпадает с нормальным распределением или с любым другим теоретическим распределением.  
  
Основные этапы алгоритма теста Колмогорова-Смирнова:  
Сбор данных. Получаем выборку, для которой нужно проверить соответствие распределению.  
  
Определение теоретического распределения. Выбираем теоретическое распределение, с которым будем сравнивать данные (например, нормальное, равномерное и т.д.).  
  
Построение эмпирической функции распределения (ЭФР):  
  
Вычисляем кумулятивные частоты значений в выборке, чтобы построить эмпирическую функцию распределения.  
  
Построение теоретической функции распределения (ТФР):  
  
На основе выбранного теоретического распределения рассчитываем его кумулятивную функцию распределения для каждого значения в выборке.  
  
Вычисление статистики Колмогорова-Смирнова:  
  
Определяем максимальное отклонение между эмпирической и теоретической функциями распределения: D = max | F\_эмп(x) - F\_теор(x) |, где F\_эмп(x) — значение эмпирической функции распределения, F\_теор(x) — значение теоретической функции распределения для каждого значения x в выборке.  
  
Сравнение с критическим значением:  
  
Полученное значение D сравнивается с критическим значением для заданного уровня значимости (обычно 0,05 или 0,01), которое зависит от объема выборки.  
Если D превышает критическое значение, гипотеза о совпадении распределений отклоняется.  
  
Интерпретация результатов:  
  
Если D меньше критического значения: гипотеза о том, что данные следуют теоретическому распределению, не отклоняется.  
Если D больше критического значения: гипотеза о соответствии распределению отклоняется, что говорит о значительных отклонениях данных от выбранного распределения.  
  
Тест Колмогорова-Смирнова часто используется для проверки нормальности и других распределений. Он также применим для двухвыборочного теста, когда нужно проверить, принадлежат ли две выборки одному и тому же распределению.

Jarque-Bera

Тест Джарка-Бера (Jarque-Bera) — это статистический тест, используемый для проверки нормальности распределения данных. Он основывается на оценке асимметрии (сместности) и эксцесса (пиковости) распределения, чтобы определить, насколько распределение данных отличается от нормального.  
  
Основные этапы алгоритма теста Джарка-Бера:  
Сбор данных. Получаем выборку, для которой нужно проверить нормальность.  
  
Вычисление параметров:  
  
n: объем выборки.  
Среднее значение выборки.  
Стандартное отклонение выборки.  
  
  
Рассчитываем асимметрию и эксцесс:  
  
Асимметрия (skewness). Измеряет, насколько данные симметричны относительно среднего. Формула: S = (1/n) \* сумма [(x\_i - среднее) / стандартное отклонение]^3.  
Эксцесс (kurtosis). Показывает, насколько распределение «пикообразно» или «плосковершинно». Формула: K = (1/n) \* сумма [(x\_i - среднее) / стандартное отклонение]^4 - 3.  
  
Расчет статистики теста Джарка-Бера: JB = (n/6) \* (S^2 + (K^2)/4). Чем больше значение JB, тем сильнее отклонение от нормальности.  
  
Сравнение с критическим значением:  
  
Полученное значение статистики JB сравнивается с критическим значением из распределения хи-квадрат с 2 степенями свободы на выбранном уровне значимости (обычно 0,05).  
  
Если JB превышает критическое значение, то гипотеза нормальности отклоняется.  
  
Интерпретация результатов:  
Если JB меньше критического значения: гипотеза о нормальности не отклоняется, и можно предположить, что данные распределены нормально.  
Если JB больше критического значения: гипотеза о нормальности отклоняется, что говорит о наличии значительной асимметрии или отклонений от нормальной формы распределения.  
  
Этот тест полезен для предварительного анализа данных и проверки предположения о нормальности, что важно во многих статистических методах и эконометрических моделях.

Статистика Jarque-Bera: 19.57625069164637

p-значение: 5.611399306311759e-05

Данные не распределены нормально

Shapiro-Wilk

Статистика Shapiro-Wilk: 0.8510308430092555

p-значение: 0.0003587078429321765

Распределение данных отличается от нормального

Helwig

Шаг 1: Сортируем данные и определяем размер выборки

Шаг 2: Оценка среднего и стандартного отклонения

Шаг 3: Вычисляем эмпирическую функцию распределения (ЭФР)

Шаг 4: Строим теоретическую нормальную функцию распределения (НФР)

Шаг 5: Вычисляем максимальное отклонение между ЭФР и НФР

Вывод результата

Максимальное отклонение (D): 0.2560889275648254

Гипотеза о нормальности отвергается на уровне значимости 0.05.

Сравнение тестов

Сравнение методов согласия Хельвига, Шапиро-Вилька и Джарка-Бера (Jarque-Bera) полезно для выбора подходящего теста для проверки нормальности распределения данных. Каждый из этих методов имеет свою область применения и особенности, которые могут быть полезны в разных контекстах.  
  
1. Тест Хельвига  
Цель: Метод Хельвига основан на анализе корреляций и используется для оценки согласия признаков, особенно в социально-экономических и психометрических исследованиях.  
Применение: Обычно применяется для оценки многомерного согласия признаков или при проведении факторного анализа.  
Преимущества:  
Хорошо подходит для многомерных данных, поскольку анализирует согласие между несколькими переменными.  
Позволяет оценить общую структуру корреляций между признаками, что важно для анализа взаимозависимости.  
Недостатки:  
Не подходит для проверки нормальности распределения данных.  
Может требовать больших выборок для корректного анализа многомерных данных.  
  
2. Тест Шапиро-Вилька  
Цель: Проверка нормальности распределения данных в выборке.  
Применение: Часто используется для малых и средних выборок (до 2000 наблюдений), чтобы оценить, насколько распределение данных близко к нормальному.  
Преимущества:  
Очень чувствителен к отклонениям от нормальности, особенно в малых выборках.  
Является одним из самых мощных тестов для проверки нормальности, так как учитывает порядок значений в выборке.  
Недостатки:  
Может давать ложные результаты для больших выборок (более 2000 наблюдений), так как становится излишне чувствительным к малейшим отклонениям.  
Не подходит для многомерных данных, так как используется для одномерного распределения.  
  
3. Тест Джарка-Бера (Jarque-Bera)  
Цель: Проверка нормальности распределения путем оценки асимметрии (skewness) и эксцесса (kurtosis).  
Применение: Часто применяется для данных больших объемов, особенно в эконометрических и финансовых исследованиях.  
Преимущества:  
Хорошо подходит для больших выборок, так как рассчитывается на основе асимметрии и эксцесса, которые более устойчивы в больших объемах данных.  
Удобен для случаев, когда нужны простые показатели нормальности (асимметрия и эксцесс).  
Недостатки:  
Менее чувствителен для малых выборок, так как асимметрия и эксцесс могут быть нестабильными.  
Не учитывает порядок значений в выборке, что делает его менее точным для малых выборок.  
  
Вывод:  
Для малых выборок (до 2000 наблюдений) тест Шапиро-Вилька наиболее подходит для проверки нормальности, поскольку он высокочувствителен к отклонениям и учитывает порядок значений.  
Для больших выборок (более 2000 наблюдений) тест Джарка-Бера предпочтителен, так как он основан на асимметрии и эксцессе, что стабильно в больших объемах данных.  
Тест Хельвига лучше использовать, когда требуется оценить согласие нескольких переменных, а не нормальность, так как он лучше подходит для анализа многомерных зависимостей.  
Таким образом, выбор метода зависит от цели исследования, объема выборки и характеристик данных.