НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Дисциплина: «Анализ данных»

Домашнее задание на тему:
«Лабораторная работа №7»

Выполнил: Осипов Лев,
студент группы 301ПИ (1).

Москва, 2015 г.

СОДЕРЖАНИЕ

ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

ЗАДАНИЕ 1

По определению ошибка – это разница между ординатой точки на графике и ординатой истинного местоположения точки. Первый график демонстрирует приближенность всех значений ошибки к нулю, что говорит о качестве и о том, что этот график соответствует методу наименьших квадратов.

Что касается второго и третьего графиков, они отличаются тем, что третий график расположен выше по оси ординат, а второй график в какой-то степени сбалансирован относительно нее. Так как при подсчете общей величины ошибок (сумма квадратов) в третьем случае это величина выйдет больше, есть основания предполагать, что МНК соответствует второй график, а третий не соответствует.

ЗАДАНИЕ 2

Первое утверждение неверно. Его можно опровергнуть следующим образом: представим, что две первые регрессии являются параллельными прямыми (разумеется, с положительным коэффициентами k), но при этом выборка, имеющая регрессию, пролегающую левее по оси абсцисс, находится выше по оси ординат, чем другая выборка. В таком случае регрессия объединения этих выборок будет прямой с отрицательным коэффициентом k.

Второе утверждение верно, так как k считается по следующей формуле:

$$k = (\sum_{i=1}^{n} x_i y_i)/(\sum_{i=1}^{n} x_i^2)$$

ЗАДАНИЕ 3

Перед обучением нормализация проводится для выбора метрики, в которой аппроксимация данных будет происходить наилучшим образом. Так как исходные данные могут быть предоставлены в различных единицах измерения, проводится нормализация (например, относительно максимального значения метрики или же по-другому, в зависимости от свойств данных).

ПРАКТИЧЕСКАЯ ЧАСТЬ

Для решения задания была написана программа, формирующая показатель состояния рынка акций с помощью PCA.

Для начала была построена гистограмма значений корреляций цен на акции:

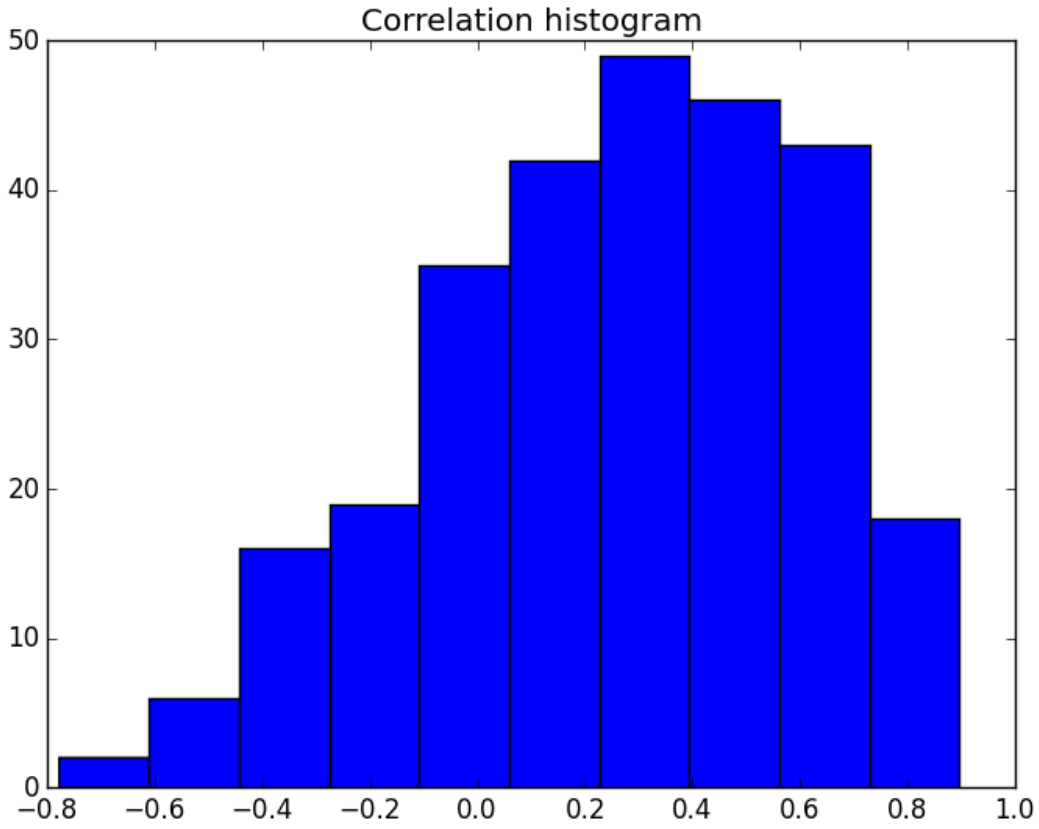

Рис. 1. Гистограмма значений корреляций цен на акции

Видно, что положительных значений наблюдается большее количество, поэтому мы можем считать, что данные сильно скоррелированы и использовать на них PCA.

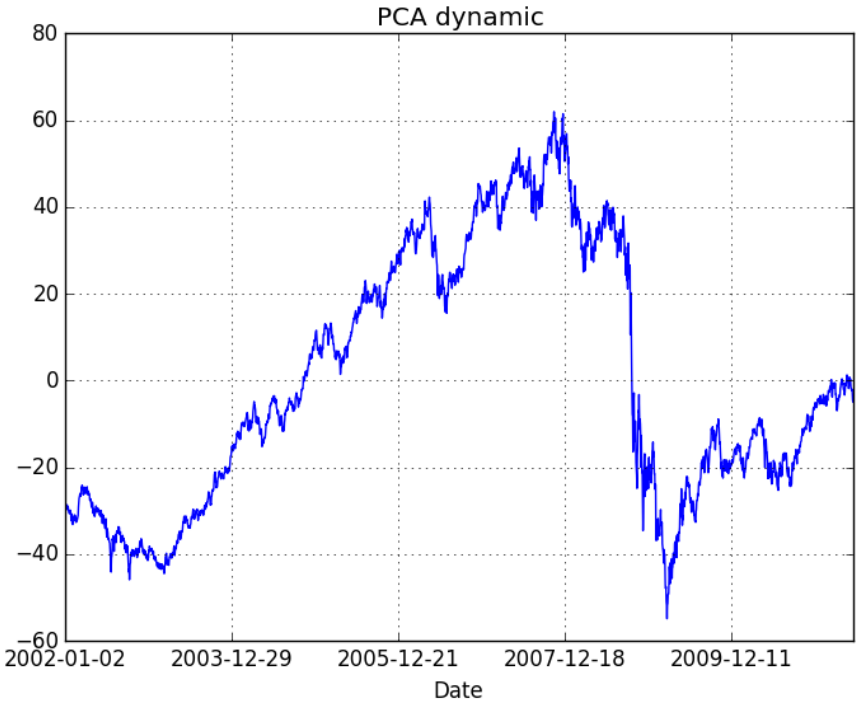

Рис. 2. Динамика показателя, полученного после PCA

```
First component 0.460008287587
```

Рис. 3. Значимость первой компоненты

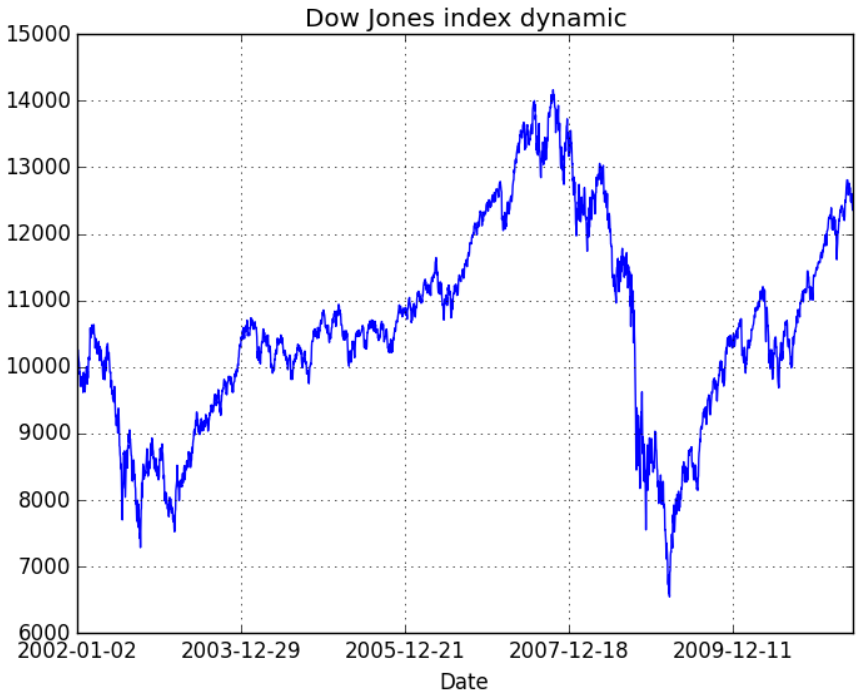

Рис. 4. Динамика реального индекса Доу-Джонса

В целом, следует отметить, что основные тренды нашего показателя и индекса Доу-Джонса совпадают.

Сильное падение нашего показателя (впрочем, как и индекса Доу-Джонса) приходится на 2008 год. Это правдоподобно, потому что в 2008 году наблюдалось начало мирового экономического кризиса.

СПИСОК ЛИТЕРАТУРЫ

1) **Анализ данных (Программная инженерия)** – http://wiki.cs.hse.ru/Анализ_данных_(Программная_инженерия)

ТЕКСТ ПРОГРАММЫ

```
__author__ = 'Lev Osipov'

import pandas as pd
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA


# Task 1
data = pd.read_csv('stock_prices.csv')

# Task 2
prices = data.iloc[:, 1:]
correlation_matrix = prices.corr()
correlations = []
for i in range(correlation_matrix.shape[0]):
    values = correlation_matrix.icol(i).values[i + 1:]
    for j in range(len(values)):
        correlations.append(values[j])
plt.hist(correlations)
plt.title("Correlation histogram")
plt.show()

# Task 3
pca = PCA(1)
reduction = pca.fit_transform(prices)
print "First component", pca.explained_variance_ratio_[0]
series = pd.Series(reduction[:, 0], data['Date'])
series.plot(title='PCA dynamic')
plt.show()

# Task 4
dji = pd.read_csv('dji.csv')
series = pd.Series(dji['Close'].values, dji['Date'])
series.plot(title='Dow Jones index dynamic')
plt.show()
```