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 часть3**

**Задание 13**

**Задание 23**

**Задание 33**

**Практическая часть4**

**Список литературы7**

**Текст программы8**

**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:



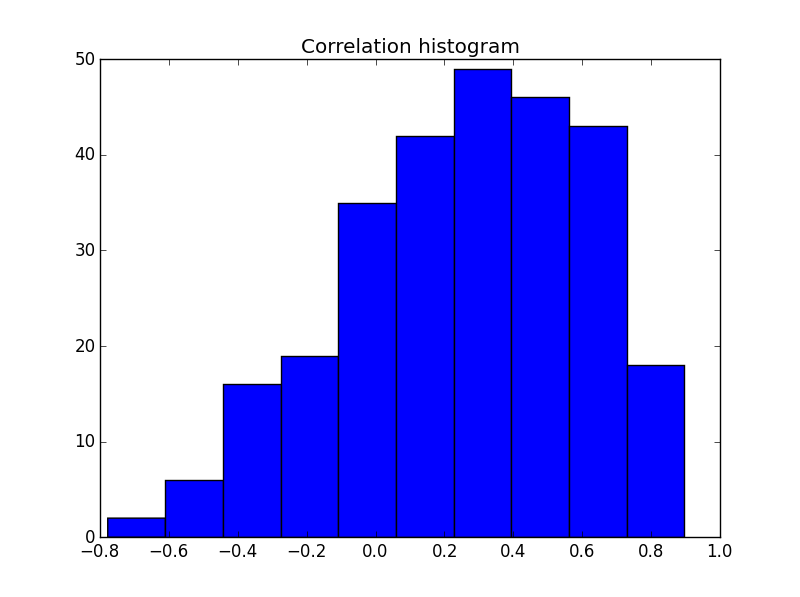
**ЗАДАНИЕ 3**

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**

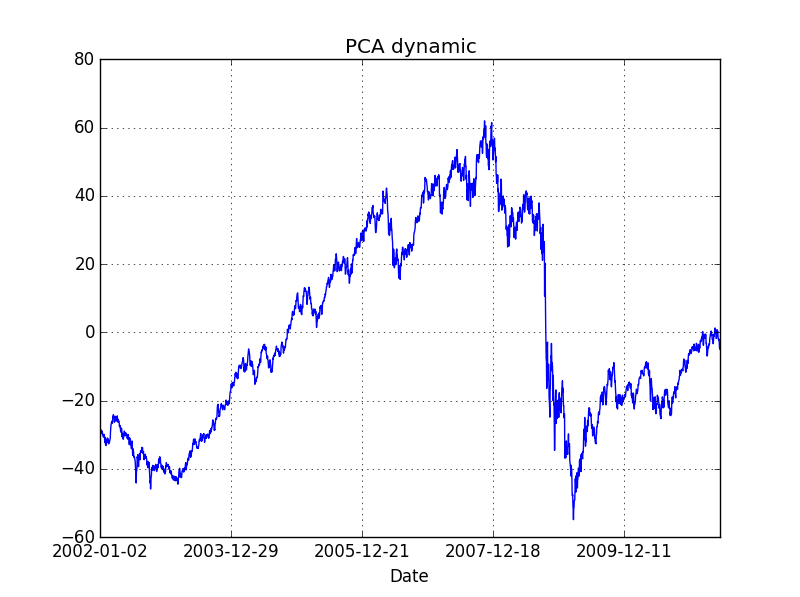
Для решения задания была написана программа, формирующая показатель состояния рынка акций с помощью PCA.

Для начала была построена гистограмма значений корреляций цен на акции:

****

*Рис. 1. Гистограмма значений корреляций цен на акции*

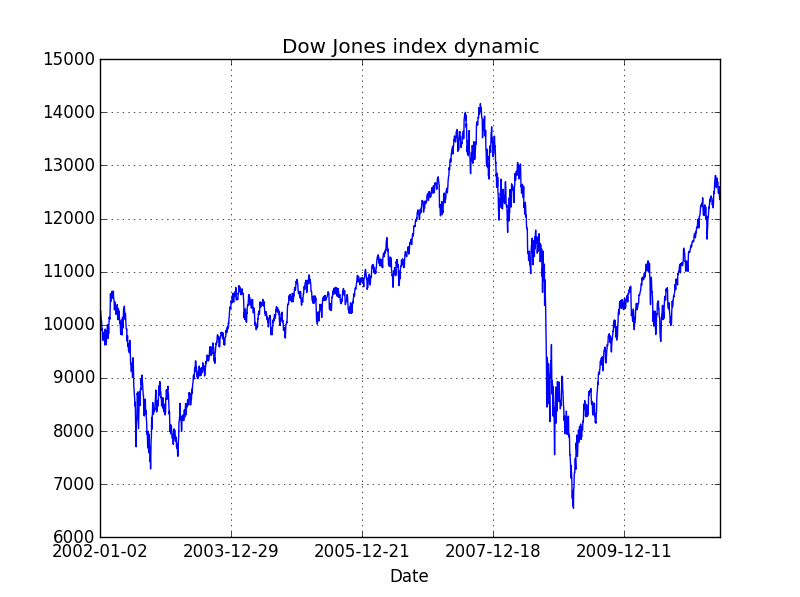
Видно, что положительных значений наблюдается большее количество, поэтому мы можем считать, что данные сильно скоррелированы и использовать на них PCA.

****

*Рис. 2. Динамика показателя, полученного после PCA*

**

*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()