Міністерство освіти і науки України
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ”

Кафедра прикладної математики

МЕТОДИЧНІ РЕКОМЕНДАЦІЇ

до виконання лабораторних робіт
з кредитного модуля "Аналіз даних"

Затверджено на засіданні
кафедри прикладної математики

Протокол № 11

від «20» червня 2018 р.

Завідувач кафедри
прикладної математики

_____ О.Р. Чертов

«____»_____ 2018 р.

Київ – 2018

Методичні рекомендації до виконання лабораторних робіт з кредитного модуля "Аналіз даних" для студентів зі спеціальності 113 Прикладна математика освітньо-кваліфікаційного рівня бакалавр за денною формою навчання складена відповідно до програми навчальної дисципліни «Аналіз даних».

Розробники методичних рекомендацій:

професор кафедри прикладної математики, д.ф.-м.н., доцент
Пашко Анатолій Олексійович

Методичні рекомендації затверджено на засіданні кафедри прикладної математики.

Протокол № 11 від «20» червня 2018 р.

Завідувач кафедри

_____ О.Р. Чертов

«____»_____ 2018 р.

Мета та завдання лабораторних робіт

Метою лабораторного практикуму є формування у студентів навичок і вмінь пов'язаних з аналізом та обробкою даних:

аналізувати дані, що використовуються в інформаційних системах;

обирати інформаційну технологію аналізу даних відповідно до визначених вимог до інформаційної системи;

забезпечувати обробку даних з використанням сучасних інформаційних технологій;

виконувати розрахунок основних характеристик вибіркових даних.

Під час викладання лекційного матеріалу з кредитного модуля «Аналіз даних» потрібно взяти до уваги наступне:

- в лекційному матеріалі слід зосередитись на теоретичних аспектах теорії аналізу прикладних даних,
- вивчення програмного забезпечення для аналізу даних на лекційних заняттях є недоцільним.

При проведенні лабораторних робіт слід акцентувати увагу студентів на прикладні аспекти аналізу даних, а саме аналіз даних для різних прикладних задач і оцінювання основних характеристик.

Основні завдання циклу лабораторних робіт: аналіз характеристик статистичних даних та їх реалізація в інформаційних системах для конкретних практичних задач.

ТЕМИ І ПЛАНИ ЛАБОРАТОРНИХ ЗАНЯТЬ

Лабораторне заняття, як одна із форм навчальних занять, розрахована на виконання студентами в електронному вигляді певної задачі з використанням ПК. Викладач організує індивідуальну роботу студентів на ПЕОМ з метою формування умінь та навичок практичного використання певних оболонок, програм тощо. На лабораторному занятті студент під керівництвом викладача проводить експерименти або дослідження з метою практичного підтвердження окремих теоретичних положень, набуває уміння роботи з сучасним програмним забезпеченням, обчислювальною технікою, вимірювальною апаратурою, оволодіває методикою експериментальних досліджень в конкретній предметній галузі та обробки отриманих результатів.

Під час виконання лабораторної роботи створюються умови для перевірки та виявлення інтелектуального рівня студентів.

Навчальні програми з переліком тем та питань дисципліни «Аналіз даних» студенти отримують на першому практичному занятті. Для самостійного опанування тем предмета студенти можуть використовувати не тільки зазначений список основної літератури, а також інші джерела інформації, можливості Internet та додаткову літературу.

На першому занятті викладач вказує на основні теми предмета для практичного засвоєння, роз'яснює загальні положення, надає рекомендації по вивченню та опануванню всіх розділів, загострює увагу на найбільш

важливих «вузлових» питаннях. На заключній лабораторній роботі розглядаються теми, які стали найбільш важкими для самостійного опанування.

Основні завдання циклу лабораторних робіт:

№ з/п	Назва та завдання лабораторної роботи	Кількість ауд. годин
1	**Організація та проведення вибіркових обстежень**. Завдання: Сплануйте власне обстеження. Сформулюйте мету обстеження, визначіть популяцію та фрейм даних. Складіть анкету вибіркового обстеження. Які величини будуть вивчатись і які параметри популяції підлягають оцінюванню. Які методи збору даних можуть бути застосовані. Проведіть обстеження і підготуйте звіт.	4
2	**Оцінювання параметрів вибіркових досліджень.** Завдання: Для заданої вибірки оцінити основні параметри. Розглянути випадки, коли вибірка отримана – випадкова вибірка без повернення, систематична вибірка, стратифікована вибірка, кластерна вибірка. Порівняти отримані результати.	4
3	**Алгоритми одно факторного і двох факторного дисперсійного аналізу**. Завдання: Провести дисперсійний аналіз даних, відповідно до варіанту, при довірчій ймовірності *a=0.99*. Провести двох факторний дисперсійний аналіз даних, відповідно до варіанту, при довірчій ймовірності *a=0.95*.	4
4	**Факторний аналіз** Завдання. Для заданих змінних оцінити основні статистичні параметри (середнє, дисперсію, побудувати гістограму, перевірити гіпотезу про закон розподілу). Побудувати кореляційну матрицю вихідних ознак. Знайти власні числа і власні вектори кореляційної матриці. Виділити основні фактори. Проаналізувати отримані дані.	4
5	**Методи кластерного аналізу.** Завдання. Розробити систему правил для розпізнавання символів українського алфавіту та реалізувати алгоритм розпізнавання.	4
6	**Регресійний аналіз.** Завдання: Результати спостереження за деякою функціональною залежністю задані таблицею (відповідно варіанту). Методом найменших квадратів знайти найкращу функціональну залежність: лінійна,	4

	поліноміальна(другого та третього порядків), логарифмічна, експоненціальна. Провести статистичний аналіз отриманих коефіцієнтів для довірчої ймовірності *a=0.95*.	
7	**Алгоритми перетворення Фур'є.** Завдання: Для реалізації деякого сигналу (відповідно до варіанту) оцінити кореляційну функцію, спектральну щільність (перетворення Фур'є). Подавити просочування енергії через бокові максимуми. Порівняти результати.	4
8	**Дослідження вейвлет – перетворення.** Завдання. Побудувати вейвлет - базис для заданого батьківського вейвлету. Порівняти результати для різних батьківських вейвлетів.	4
9	**Алгоритми вейвлет – аналізу.** Завдання. Задано сигнал заданої довжини. Обчислити нелінійну апроксимацію сигналу за допомогою вейвлет – коефіцієнтів Хаара. Знайти результуючу похибку апроксимації. Прокоментувати результати.	4

Методичні рекомендації з виконання лабораторних робіт

Лабораторна робота 1. Організація та проведення вибіркових обстежень.

Мета роботи – здобути практичні навички організації та проведення вибіркових досліджень.

Завдання. Сплануйте власне обстеження. Сформулюйте мету обстеження, визначіть популяцію та фрейм даних. Складіть анкету вибіркового обстеження. Які величини будуть вивчатись і які параметри популяції підлягають оцінюванню. Які методи збору даних можуть бути застосовані. Проведіть обстеження і підготуйте звіт. В звіті треба відобразити аналіз помилок, що виникали при проведенні обстежень.

Методичні вказівки до організації та проведення вибіркового обстеження.

В якості популяції для дослідження виберіть:

- м. Київ
- студентська спільнота м. Києва
- студентська спільнота НТУУ КПІ
- студентська спільнота ФПМ
- студентська спільнота 3 курсу ФПМ.

Сплануйте вибіркове обстеження для дослідження параметрів:

- середній зріст;
- підтримка європейського вибору України (в середньому і всього);
- вболівальники "Динамо" Київ (в середньому і всього);

- індекс IQ;
- час знаходження в мережі Інтернет.

Оформіть звіт, в якому відобразіть:
- ціль вибіркового обстеження
- визначення популяції
- планування вибірки
- побудова вибіркової схеми
- одержання вибірки для обстеження
- зібрані дані
- результати обробки даних
- точність результатів.

Лабораторна робота 2. Оцінювання параметрів вибіркових досліджень.

Мета роботи – здобути практичні навички обробки результатів вибіркових досліджень скінчених популяцій.

Завдання. Для заданої вибірки оцінити основні параметри популяції. Розглянути випадки, коли вибірка – випадкова вибірка без повернення, систематична вибірка, стратифікована вибірка. Розробити програмне забезпечення. Порівняти отримані результати. Оформити звіт.

Методичні вказівки до виконання.

Довжина популяції N. Деякий параметр популяції $\{y_i\}, i = 1,2...N$ змоделюйте як рівномірний розподіл на відрізку $[10+M, MG+2M]$, M – номер студента в списку групи, MG - номер групи.

Вибрати число M - об'єм вибірки.

Простим випадковим вибором сформуйте вибірку $\{x_i\}, i = 1,2...M$ із популяції.

А). Обчисліть середнє значення і дисперсію для популяції

$$Y = \sum_{i=1}^{N} y_i, \qquad Ys = \frac{1}{N}\sum_{i=1}^{N} y_i, \qquad DY^2 = \frac{\sum_{i=1}^{N}(y_i - Ys)^2}{N-1}.$$

Б). Оцініть суму і середнє значення для популяції за вибіркою

$$\hat{Y} = N\sum_{i=1}^{M}\frac{x_i}{M}, \qquad \hat{Ys} = \frac{1}{M}\sum_{i=1}^{M} x_i, \qquad S^2 = \frac{\sum_{i=1}^{M}(x_i - \hat{Ys})^2}{M-1}.$$

Дисперсія оцінок суми і середнього в залежності від об'єму випадкової вибірки:

$$D(\hat{Ys}) = \frac{S^2}{M}\left(1 - \frac{M}{N}\right) \text{ та } D(\hat{Y}) = \frac{S^2N^2}{M}\left(1 - \frac{M}{N}\right)$$

Верхній і нижній $100(1-\alpha)\%$ довірчі інтервали для середнього значення і загальної суми популяції відповідно дорівнюють

$$\hat{Ys} \pm \frac{z_{1-\frac{\alpha}{2}} S}{\sqrt{M}}\sqrt{1-\frac{M}{N}} \quad \text{та} \quad N\hat{Ys} \pm \frac{z_{1-\frac{\alpha}{2}} SN}{\sqrt{M}}\sqrt{1-\frac{M}{N}}$$

де z є відповідним 1-*a/2* квантилем нормального розподілу $N(0,1)$. Найбільш часто використовуються наступні квантилі:

Рівень довіри (%)	50	80	90	95	99
z	0,67	1,28	1,64	1,96	2,58

В). Написати програму для обчислення основних показників вибіркового обстеження. Провести розрахунки для

$N = 10000$, $M = 100$ та $M = 1000$.

Порівняти результати.

Г). Сформувати систематичну вибірку таких же розмірів. Порівняти отримані результати.

Зауваження. Наведені формули для довірчих інтервалів використовуються при розмірах вибірок $M > 50$, при менших розмірах величину z беруть з таблиць розподілу Ст'юдента з (n-1) степенями свободи.

Д). Провести аналогічні розрахунки для стратифікованої вибірки.

Використаємо наступні позначення:

y_{hi} - значення i–го елемента зі страти h,

$w_h = \frac{N_h}{n_h}$ - вибіркова вага,

$\bar{y}_h = \frac{\sum_{i=1}^{N_h} y_{hi}}{N_h}$ - середнє значення для страти h,

$Y_h = \sum_{i=1}^{N_h} y_{hi}$ - загальна сума для страти h,

$Y = \sum_{h=1}^{H} Y_h$ - загальна сума для всієї популяції,

$\bar{y}_U = \frac{Y}{N} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{N_h} y_{hi}}{N}$ - середнє значення для всієї популяції,

$S_h^2 = \frac{\sum_{i=1}^{N_h} (y_{hi} - \bar{y}_h)^2}{N_h - 1}$ - дисперсія в страті h,

$W_h = \frac{N_h}{N}$ - вага страти h.

$\hat{\bar{y}}_h = \frac{\sum_{i \in b_h} y_{hi}}{n_h}$ - оцінка середнього для страти *h, h=1,...,N*

$\hat{Y}_h = \frac{N_h}{n_h}\sum_{i \in b_h} y_{hi}$ - оцінка загальної суми для страти *h, h=1,...,N*

$s_h^2 = \frac{\sum_{i \in b_h} (y_{hi} - \hat{\bar{y}}_h)^2}{n_h - 1}$ - оцінка вибіркової дисперсії для страти *h, h=1,...,N.*

Оцінки суми та середнього популяції при використанні стратифікованої вибірки будуть відповідно мати вигляд

$$\hat{Y}_{st} = \sum_{h=1}^{H} \hat{Y}_h = \sum_{h=1}^{H} N_h \hat{\bar{y}}_h ,$$

$$\hat{\bar{y}}_{st} = \frac{\hat{Y}_{st}}{N} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{\bar{y}}_h = \sum_{h=1}^{H} W_h \hat{\bar{y}}_h .$$

З останньої формули видно, що середнє значення популяції є зваженим середнім від середніх для кожної страти з вагами, які дорівнюють відносним розмірам страт.

Незміщена оцінка для дисперсії:

$$\hat{D}(\hat{Y}_{st}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{s_h^2}{n_h}$$

$$\hat{D}(\hat{\bar{y}}_{st}) = \frac{1}{N^2} D(\hat{Y}_{st}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h}$$

100(1-α)% довірчий інтервал для оцінки середнього значення визначається за формулою $\hat{\bar{y}}_{st} \pm z_{\frac{\alpha}{2}} \sqrt{\hat{D}(\hat{\bar{y}}_{st})}$, де $z_{\frac{\alpha}{2}}$ є відповідним квантилем стандартного нормального розподілу $N(0,1)$.

Лабораторна робота 3. Алгоритми однофакторного і двох факторного дисперсійного аналізу.

Мета роботи – здобути практичні навички проведення і аналізу даних однофакторного та двохфакторного дисперсійного аналізу.

Завдання. Провести дисперсійний аналіз даних, відповідно до варіанту, при довірчій ймовірності a=0.95.

Провести двох факторний дисперсійний аналіз даних, відповідно до варіанту, при довірчій ймовірності a=0.95.

За результатами оформити звіт. В звіті відобразити особливості реалізації алгоритмів в середовищі R Studio.

Методичні вказівки до виконання.

Номер студента в списку групи N.

1. Однофакторний аналіз

Таблиця однофакторного експерименту

n	Рівні фактора A (i)				
(j)	A_1	A_2	A_3	A_4	A_5
1000	N+rnd(1)	0.5*N+rnd(1)	0.8*N+rnd(1)	1.4*N+rnd(1)	2*N+rnd(1)

Для кожного фактору знаходимо

$$S_i^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2 = \frac{1}{n-1} \left[\sum_{j=1}^{n} x_{ij}^2 - \frac{1}{n} \left(\sum_{j=1}^{n} x_{ij} \right)^2 \right].$$

За припущенням дисперсійного аналізу - повинна мати місце рівність дисперсій. Перевірити рівність дисперсій за критерієм порівняння.

$$g = \frac{\max\limits_{1 \leqslant i \leqslant k} s_i^2}{\sum\limits_{i=1}^{k} s_i^2}.$$

Критерій порівняння подується за формулою

При $g > g_\alpha(k, n)$ нульова гіпотеза про рівність дисперсій відхиляється.

Значення статистики $g_\alpha(k, n)$ для α=0,95

k	n					
	8	9	10	11	17	37
2	0,899	0,882	0,867	0,854	0,795	0,707
3	0,733	0,711	0,691	0,673	0,606	0,515
4	0,613	0,590	0,570	0,554	0,488	0,406
5	0,526	0,504	0,485	0,470	0,409	0,335
6	0,461	0,440	0,423	0,408	0,353	0,286
7	0,410	0,391	0,375	0,362	0,310	0,249
8	0,370	0,352	0,337	0,325	0,278	0,221
9	0,338	0,321	0,307	0,295	0,251	0,199
10	0,311	0,294	0,281	0,270	0,230	0,181
12	0,268	0,253	0,242	0,232	0,196	0,153
15	0,223	0,210	0,200	0,192	0,161	0,125
20	0,175	0,165	0,157	0,150	0,125	0,096
30	0,123	0,116	0,110	0,105	0,087	0,066

При виконанні припущення про рівність дисперсій, знаходимо оцінку дисперсії, що характеризує розсіювання поза впливом фактора,

$$S_0^2 = \frac{1}{k}\sum_{i=1}^{k} S_i^2 \frac{1}{k(n-1)} \sum_{i=1}^{k}\sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2 = \frac{1}{k(n-1)} \left[\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2 - \frac{1}{n}\sum_{i=1}^{k} \left(\sum_{j=1}^{n} x_{ij} \right)^2 \right]$$

Знаходимо вибіркову дисперсію всіх спостережень

$$S^2 = \frac{1}{kn-1} \left[\sum_{i=1}^{k}\sum_{j=1}^{k} x_{ij}^2 - \frac{1}{kn} \left(\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij} \right)^2 \right]$$

Знаходимо оцінку дисперсії, що характеризує зміни параметра, пов'язані з фактором

$$S_A^2 = \frac{n}{k-1} \sum_{i=1}^{k} (\bar{x}_i - \bar{\bar{x}})^2 . \qquad \bar{\bar{x}} = \frac{1}{k}\sum_{i=1}^{k} \bar{x}_i; \quad \bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij}.$$

Оцінка впливу фактора на зміни середнього значення визначається відношенням (вплив значущий з ймовірністю 1-α)

$$\frac{S_A^2}{S_0^2} > F_\alpha[k-1; k\,(n-1)]$$

де $F_\alpha(f_1,f_2)$ - α- квантиль F- розподілу з f_1 та f_2 степенями свободи.

2. Двохфакторний аналіз

Таблиця двохфакторного експерименту

Рівні фактора B(j)	Рівні фактора A (i)				
	A_1	A_2	A_3	A_4	A_5

B_1	N+rnd(1)	3.5*N+rnd(1)	3.8*N+rnd(1)	1.4*N+rnd(1)	2*N+rnd(1)
B_2	N+rnd(1)	2.5*N+rnd(1)	2.8*N+rnd(1)	2.4*N+rnd(1)	3*N+rnd(1)
B_3	N+rnd(1)	1.5*N+rnd(1)	1.8*N+rnd(1)	3.4*N+rnd(1)	4*N+rnd(1)
B_4	N+rnd(1)	0.5*N+rnd(1)	0.8*N+rnd(1)	4.4*N+rnd(1)	5*N+rnd(1)

В кожній клітинці таблиці зберігається масив із n=100 значень.

Якщо фактори А і В незалежні:

Знаходимо середнє значення в кожній клітинці x_{ij}.

Обчислюємо основні показники

$$Q_1 = \sum_{i=1}^{k}\sum_{j=1}^{m} x_{ij}^2; \quad Q_2 = \frac{1}{m}\sum_{i=1}^{k} X_i^2; \quad Q_3 = \frac{1}{k}\sum_{j=1}^{m} X_{j'}^2;$$

$$Q_4 = \frac{1}{mk}\left(\sum_{i=1}^{k} X_i\right)^2 = \frac{1}{mk}\left(\sum_{j=1}^{m} X_{j'}\right)^2.$$

X_i - сума по стовпчиках (рівень фактора A_i), $X_{j'}$ - сума по рядках (рівень фактора B_j).

Знаходимо оцінки дисперсій

$$S_0^2 = \frac{Q_1 + Q_4 - Q_2 - Q_3}{(k-1)(m-1)}; \quad S_A^2 = \frac{Q_2 - Q_4}{k-1}; \quad S_B^2 = \frac{Q_3 - Q_4}{m-1}.$$

Якщо $\frac{S_A^2}{S_0^2} > F_\alpha(f_1, f_2)$ для f_1=k-1, f_2=(k-1)(m-1) то фактор А є значущим. Аналогічно для фактора В при f_1=m-1, f_2=(k-1)(m-1) .

Якщо фактори А і В залежні:

Знаходимо додатково

$$Q_5 = \sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{\nu=1}^{n} x_{ij\nu}^2. \quad S_{AB}^2 = \frac{Q_5 - nQ_1}{mk(n-1)},$$

взаємодію факторів перевіряємо за критерієм

$\frac{nS_0^2}{S_{AB}^2} > F_\alpha(f_1, f_2)$, де $f_1 = (k-1)(m-1)$, $f_2 = mk(n-1)$.

Лабораторна робота 4. Факторний аналіз.

Мета роботи – здобути практичні навички проведення факторного аналізу даних.

Завдання. Для заданих змінних оцінити основні статистичні параметри (середнє, дисперсію, побудувати гістограму, перевірити гіпотезу про закон розподілу). Побудувати кореляційну матрицю вихідних ознак. Знайти власні числа і власні вектори кореляційної матриці. Виділити основні фактори. Проаналізувати отримані дані.

Методичні вказівки до виконання. Алгоритм головних компонентів.

Крок 1. Нормалізація всіх пояснювальних змінних:

$$x_{ij}^{*} = \frac{x_{ij} - \bar{x}_j}{\sigma_{x_j}}, \quad i = \overline{1,n};\ j = \overline{1,m}.$$

Крок 2. Обчислення кореляційної матриці

$$r = \frac{1}{n}\left(X^{*\prime}\, X^{*}\right).$$

Крок 3. Знаходження характеристичних чисел матриці r з рівняння

$$\left|r - \lambda E\right| = 0, k = \overline{1,m},$$

де E — одинична матриця розміром $m \times m$.

Крок 4. Власні значення λ_k упорядковуються за абсолютним рівнем вкладу кожного головного компонента до загальної дисперсії.

Крок 5. Обчислення власних векторів a_k розв'язуванням системи рівнянь

$$\left(r - \lambda E\right)a = 0$$

за таких умов:

$$a_j' a_k = \begin{cases} 0\,(j \neq k), \\ 1\,(j = k). \end{cases}$$

Крок 6. Знаходження головних компонентів — векторів

$$z_k = x \cdot a_k,\ k = \overline{1,m}.$$

Головні компоненти мають задовольняти умови:

$$\sum_{i=1}^{n} z_{k,i} = 0,\ i = \overline{1,n};$$

$$\frac{1}{n} z_k' z_k = \lambda_k,\ k = \overline{1,m};$$

$$z_j' z_k = 0,\ j = \overline{1,m},\ j \neq k.$$

Крок 7. Визначення параметрів моделі $\hat{Y} = Z\hat{b}$:

$$\hat{b} = Z^{-1} Y.$$

Крок 8. Знаходження параметрів моделі $\hat{Y} = X\hat{\beta}$:

$$\hat{\beta} = a \cdot \hat{b}.$$

Лабораторна робота 5. Методи кластерного аналізу.

Мета роботи – здобути практичні навички проведення кластерного аналізу даних.

Завдання. Розробити систему правил для розпізнавання символів українського алфавіту та реалізувати алгоритм розпізнавання.

Методичні вказівки до виконання.

Букви отримані в результаті сканування тексту. Букви українського алфавіту задаються матрицею розміром $m * n$. Замальована клітинка має значення 1, а біла – 0.

Дана база букв українського алфавіту.

Є відсканований текст. Розпізнати його.

Для розпізнавання букв використати відстані:
евклідову

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(x_{ki} - x_{kj}\right)^2},$$

Мінковського

$$d_{ij} = \sqrt[r]{\sum_{k=1}^{p} \left(x_{ki} - x_{kj}\right)^r},$$

Хеммінгову

$$d_{ij} = \sum_{k=1}^{p} \left|x_{ki} - x_{kj}\right|,$$

Чебишева

$$d_{ij} = \sup_{1 \le k \le p} \left|x_{ki} - x_{kj}\right|.$$

Порівняти та прокоментувати отримані результати.

Лабораторна робота 6. Регресійний аналіз.

Мета роботи – здобути практичні навички проведення регресійного аналізу даних.

Завдання. Результати спостереження за деякою функціональною залежністю задані таблицею (відповідно варіанту). Методом найменших квадратів знайти найкращу функціональну залежність: лінійна, поліноміальна(другого та третього порядків).

Провести статистичний аналіз отриманих коефіцієнтів для довірчої ймовірності a=0.95.

Методичні вказівки до виконання.

А) Задано: $n=1000,\ \delta=0.95$

$x_i = i + (rnd(1) * N)/NG$

$y_i = N * rnd(1) * x_i + NG * rnd(1) + N, \quad i = 1,2,\dots,n$

N- номер студента в списку групи,

NG- номер групи.

Знайти оцінки параметрів лінійної регресії $y = \alpha + \beta x$ методом найменших квадратів.

Перевірити наявність викидів у регресії - якщо $R > R_\delta$, то y_i, що відповідає максимальному значенню відношення $\frac{e_i}{S_i}$ є викидом з достовірністю δ, де $\hat{y}_i = a + bx_i$

$$e_i = y_i - \hat{y}_i,$$

$$R = \max\left|\frac{e_i}{S_i}\right|$$

$$S_i^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}\left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

Значення $R_\delta \approx 4$.

Якщо є викиди, то їх видалити і провести оцінки спочатку.

Перевірити гіпотези про відповідність оцінок коефіцієнтів істинним значенням та адекватність моделі.

1. $H_0: \beta = b$. Значення коефіцієнта є значимим з достовірністю δ, якщо $|b| > t_{\frac{1+\delta}{2}} S_\beta$,

де $t_{\frac{1+\delta}{2}}$ – коефіцієнт розподілу Стюдента з (*n-2*) степенями свободи,

$$S_\beta = \frac{S}{S_x\sqrt{n-1}}; \quad S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - a - bx_i)^2;$$

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2; \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i;$$

2. $H_0: \alpha = a$. Значення коефіцієнта є значимим з достовірністю δ, якщо $|a| > t_{\frac{1+\delta}{2}} S_\alpha$,

де $t_{\frac{1+\delta}{2}}$ – коефіцієнт розподілу Стюдента з (*n-2*) степенями свободи,

$$S_\alpha = S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)S_x^2}},$$

3. H_0: **модель адекватна.** Модель адекватна з достовірністю δ, якщо $\frac{S^2}{S_y^2} < F_\delta$,

де F_δ – квантиль розподілу Фішера з (*n-2*) та (*n-1*) степенями свободи.

В) Методом найменших квадратів знайти найкращу функціональну залежність: лінійна, поліноміальна(другого та третього порядків). Для вибору найкращої скористатись значенням S^2.

Результати оформити у вигляді звіту.

Лабораторна робота 7. Алгоритми перетворення Фур'є.

Мета роботи – здобути практичні навички проведення спектрального аналізу даних.

Завдання. Для заданої реалізації деякого сигналу (відповідно до варіанту) оцінити коваріаційну функцію, спектральну щільність (перетворення Фур'є). Використати вагові вікна Хеммінга і Блекмана. Порівняти результати.

Методичні вказівки до виконання.

А) Задано: *n=NG/2,*

$$s2_i := 2\cdot md(1) + NG\cdot\cos\left(2M\pi\cdot\frac{i}{N}\right)\cdot(1 + 0.1\,md(1)) + 17\cdot\cos\left(\frac{4M\pi i}{N} + md(1)\right) + 3\cos\left(\frac{5M\pi i}{N}\right)\cdot(md(1) + NG)$$

$N=2^n$,

M - номер студента в списку групи,

NG- номер групи.

Провести нормування, оцінити кореляційну функцію.

Виконати перетворення Фур'є за формулами

$$A_0 := \frac{1}{N}\cdot\sum_{i=0}^{N-1}\left(s1_i\cdot\cos\left(\frac{2\pi\cdot i\cdot 0}{N}\right)\right) \qquad A_{\frac{N}{2}} := \frac{1}{N}\cdot\sum_{i=0}^{N-1}\left(s1_i\cdot\cos\left(\frac{\pi\cdot i}{1}\right)\right)$$

$$A_l := \frac{2}{N}\cdot\sum_{i=0}^{N-1}\left(s1_i\cdot\cos\left(\frac{2\pi\cdot i\cdot l}{N}\right)\right) \qquad l := 1,2..\frac{N}{2} - 1$$

$$B_j := \frac{2}{N}\cdot\sum_{i=0}^{N-1}\left(s1_i\cdot\sin\left(\frac{2\pi\cdot i\cdot j}{N}\right)\right) \qquad j := 0,1..\frac{N}{2}$$

$$C_j := \sqrt{\left(A_j\right)^2 + \left(B_j\right)^2} \qquad j := 0,1..\frac{N}{2}$$

Знайти спектр .

Побудувати графік. Обчислити частоту першої синусоїди (або крок по частоті).

Розрахунки виконати для варіантів

$s1_i := s2_i$,

$s1_i := s2_i\cdot va_i$ $\qquad va_i := 0.42 - 0.5\cos\left(\frac{2\pi i}{N}\right) + 0.08\cos\left(\frac{4\pi i}{N}\right)$, ,

$s1_i := s2_i\cdot vb_i$ $\qquad vb_i := 0.54 - 0.46\cdot\cos\left(\frac{2\pi i}{N}\right)$, .

Порівняти отримані результати.

Виконати обернене перетворення Фур'є

$$d1_i := \sum_{j=0}^{\frac{N}{2}}\left(A_j\cdot\cos\left(\frac{2\pi j\cdot i}{N}\right)\right) + \sum_{j=0}^{\frac{N}{2}}\left(B_j\cdot\sin\left(\frac{2\pi j\cdot i}{N}\right)\right)$$

.

Порівняти початковий і результуючий масиви. Співпали?

*) Запрограмувати швидке перетворення Фур'є (або використати готову програму). Провести аналогічні розрахунки. Порівняти отримані результати.

Лабораторна робота 8. Дослідження вейвлет – перетворення.

Мета роботи – здобути практичні навички побудови вейвлет – базису для заданого батьківського вейвлету.

Завдання. Побудувати вейвлет - базис для заданого батьківського вейвлету. Порівняти результати для різних батьківських вейвлетів.

Методичні вказівки до виконання.

Для батьківських вейвлетів Хаара, Шенона, Гауссового побудувати вейвлет - базис. Представити базис графічно.

Лабораторна робота 9. Алгоритми вейвлет – аналізу.

Мета роботи – здобути практичні навички проведення вейвлет - аналізу даних.

Завдання. Задано сигнал заданої довжини. Обчислити нелінійну апроксимацію сигналу за допомогою вейвлет – коефіцієнтів Хаара. Знайти результуючу похибку апроксимації.

В звіті прокоментувати отримані результати.

Методичні вказівки до виконання.

Задано: $n=NG/2$,

$$s1_i := 2\cdot md(1) + NG\cdot\cos\left(2M\pi\cdot\frac{i}{N}\right)\cdot(1 + 0.1\,md(1)) + 17\cdot\cos\left(\frac{4M\pi i}{N} + md(1)\right) + 3\cos\left(\frac{7M\pi i}{N}\right)\cdot(md(1) + NG)$$

$N=2^n$,

M - номер студента в списку групи,

NG- номер групи.

Виконати пряме і обернене вейвлет - перетворення для кожного базису.

$M1=N$,

$$f2(j,k,x) := 2^{\frac{j}{2}}\cdot g1\left(2^j\cdot x - k\right)$$

, $g1(x)$ - батьківський вейвлет,

$$W(1,j) := \sum_{i=0}^{N-1}\left(s1_i\cdot f2(1,j,i)\right)$$

,

$$d_i := \sum_{l=0}^{M}\sum_{j=0}^{M1}\left(W(1,j)\,\frac{f2(1,j,i)}{2^{2l}}\right)$$

.

Порівняти початковий і відновлений масиви.

Рекомендована література

Базова література

1. Пархоменко В.М. Методи вибіркових обстежень / В.М. Пархоменко. – Київ. – 2001. – 148 с.
2. Василик О.І. Лекції з теорії і методів вибіркових обстежень / О.І. Василик, Т.О. Яковенко. - Київ: ВПЦ "Київський університет", 2010. - 208 с.
3. Джессен Р.Д. Методы статистических обследований / Р.Д. Джессен. - М.: Финансы и статистика. - 1985.
4. Бахрушин В.Є. Методи аналізу даних : навчальний посібник для студентів / В.Є. Бахрушин. – Запоріжжя : КПУ, 2011. – 268 с.
5. Айвазян С. А. Прикладная статистика: Исследование зависимостей: Справ. изд. / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. — М.: Финансы и статистика. - 1985. — 487 с.
6. Айвазян С. А. Прикладная статистика: Классификация и снижение размерности: Справ. изд. / С. А. Айвазян, В.М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин. — М.: Финансы и статистика. - 1989. — 607 с.
7. Добеши И. Десять лекций по вейвлетам / И. Добеши. – Москва-Ижевск: НИЦ «Регулярная и хаотическая динамика». – 2004. – 464 с.
8. Бендат Дж. Прикладной анализ случайных данных / Дж. Бендат, А. Пирсол. М.: Мир. - 1989. – 540 с.
9. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников / А.И. Кобзарь. – М.: ФИЗМАТЛИТ. - 2006. – 816 с.
10. Джонсон Н. Статистика и планирование эксперимента в технике и науке. Методы обработки данных / Н. Джонсон, Ф. Аннон. – М.: Мир. – 1980. - 610 с.
11. Малла С. Вейвлеты в обработке сигналов / С. Малла. – М.: Мир.- 2005. – 672 с.
12. Майборода Р.Є. Регресія: Лінійні моделі: Навчальний посібник / Р.Є. Майборода. – К.:ВПЦ «Київський університет». - 2007. – 296 с.
13. Ugarte M.D. Probability and statistics with R / M.D. Ugarte, A.F. Militino, A.T. Arnholt. – Boca Raton, London, New York: CRC Press, Taylor&Francis Group. - 2008. – 700 p.

Допоміжна література

14. Гнеденко Б.В. Курс теории вероятностей / Б.В. Гнеденко. - М.: Наука. - 1988.
15. Закс Л. Статистическое оценивание / Л. Закс. –М.: СТАТИСТИКА. – 1976. – 598 с.
16. Шеффе Г. Дисперсионный анализ / Г. Шеффе. – М.: Наука. – 1980. – 512 с.
17. Закс Ш. Теория статистических выводов / Ш. Закс. – М.: Мир. – 1975. – 776 с.
18. Кокрен У. Методы выборочного исследования / У. Кокрен. – М.: Финансы и статистика. – 1976.

19. Блаттер К. Вейвлет – анализ. Основы теории / К. Блаттер. – М.: Техносфера. – 2004. 276 с.
20. Бахтин В.И. Введение в прикладную статистику / В.И. Бахтин. – Минск: БГУ. – 2011. – 91 с.
21. Шитиков В.К., Классификация, регрессия и другие алгоритмы Data Mining с использованием R / В.К. Шитиков, С.Э. Мастицкий. − Электронная книга, адрес доступа: https://github.com/ranalytics/data-mining. - 2017. – 351 с.
22. Спиридонов А.А. Планирование эксперимента при исследовании технологических процессов / А.А. Спиридонов. – М.: Машиностроение. – 1981. – 184 с.
23. Столниц Э. Вейвлеты в компьютерной графике. Теория и приложения / Э. Столниц, Т. ДеРоуз, Д. Салезин. – Ижевск: НИЦ «Регулярная и хаотическая динамика». – 2002. – 272 с.
24. Чуи К. Введение в вэйвлеты / К. Чуи. – М.: Мир. – 2001. – 412 с.
25. Новиков Л.В. Основы вейвлет-анализа сигналов / Л.В. Новиков. – С-Пб.: ООО»МОДУС». – 1999. – 152 с.
26. Лайонс Р. Цифровая обработка сигналов. / Р. Лайонс. - М.: ООО «Бином-Пресс». - 2006. - 656с.
27. Сергиенко А.Б. Цифровая обработка сигналов / А.Б. Сергиенко. - С-Пб.: ООО «ПитерПринт». - 2002. - 605с.

$\alpha = 0{,}05$

10	12	15	20	24	30	40	60	120	∞
242	244	246	248	249	250	251	252	253	254
19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5	19,5
8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
2,20	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
2,18	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
1,83	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00