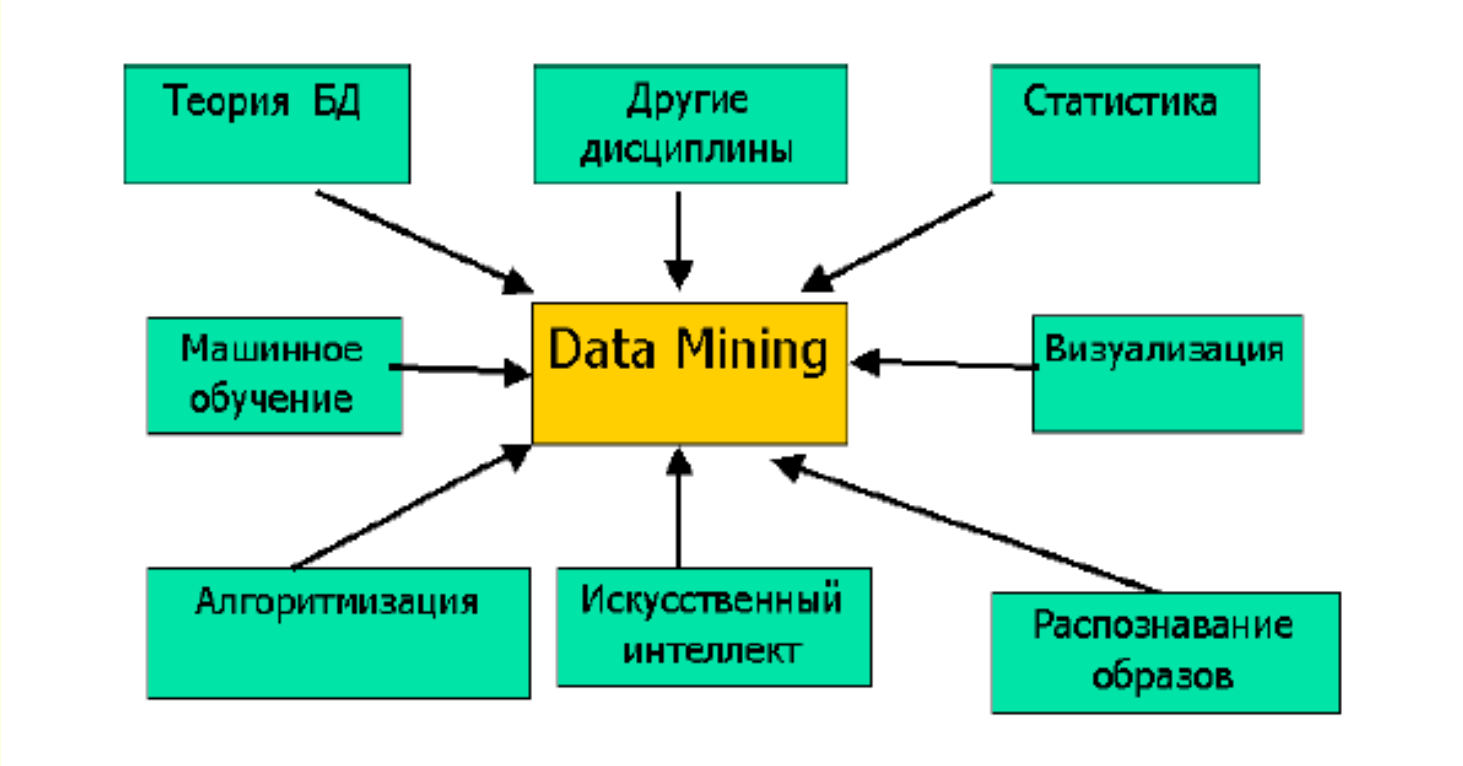
Интеллектуальный анализ данных

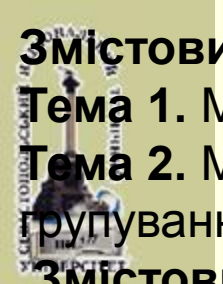


1. Сравнение статистики, машинного обучения и Data Mining

- Статистика
  - Более, чем Data Mining, базируется на теории.
  - Более сосредотачивается на проверке гипотез.
- Машинное обучение
  - Более эвристично.
  - Концентрируется на улучшении работы агентов обучения.
- Data Mining.
  - Интеграция теории и эвристик.
  - Сконцентрирована на едином процессе анализа данных, включает очистку данных, обучение, интеграцию и визуализацию результатов.

Шифр навчальної дисципліни	Назва навчальної дисципліни	Назва блока змістового модуля	Шифр блока змістових модулів	Назва теми (змістового модуля)	Шифр теми (змістового модуля)
3.04	Організація баз даних та знань	Моделювання даних	3.04.01	Системи баз даних. Основні поняття й архітектура	3.04.01.01
				Моделі даних	3.04.01.02
				Реляційна модель даних	3.04.01.03
				Теорія нормалізації реляційної моделі даних	3.04.01.04
		Мови запитів	3.04.02	Мова SQL	3.04.02.01
				Мова QBE	3.04.02.02
		Проектування та захист баз даних	3.04.03	Проектування баз даних	3.04.03.01
				Цілісність даних	3.04.03.02
				Захист баз даних	3.04.03.03
				Навігаційна обробка даних	3.04.03.04
		Класифікація баз даних	3.04.04	Розподілені бази даних	3.04.04.01
				Паралельні бази даних	3.04.04.02
				Дедуктивні бази даних	3.04.04.03
				Об'єктно-орієнтовані бази даних	3.04.04.04
				Бази даних в Інтернеті	3.04.04.05
				Бази знань	3.04.04.06
3.05	Інтелектуальний аналіз даних	Основи інтелектуального аналізу даних	3.05.01	Методи первісної обробки даних	3.05.01.01
				Методи дослідження структури даних: візуалізація та автоматичне групування даних	3.05.01.02
		Методи використання навчальної інформації	3.05.02	Кореляційний і регресійний аналіз даних. Множинний регресійний аналіз.	3.05.02.01
				Лінійна множинна регресійна модель. Перевірка адекватності моделі	3.05.02.02
				Нелінійне оцінювання параметрів	3.05.02.03
		Методи багатомірного розвідувального аналізу	3.05.03	Кластерний аналіз. Ієрархічна та секційна кластеризації	3.05.03.01
				Методи кластеризації: процедура Мак-Кина, метод k-методів, сітчасті методи	3.05.03.02
				Растрова кластеризація об'єктів	3.05.03.03
				Лінійний дискримінантний аналіз. Побудова канонічних та класифікаційних функцій	3.05.03.04
		Методи класифікації та прогнозування	3.05.04	Дерева рішень	3.05.04.01
				Методи опорних векторів, «найближчого сусіда», Байеса	3.05.04.02
				Аналіз багатомірних угруповань	3.05.04.03
				Статистична обробка тимчасових рядів і прогнозування	3.05.04.04
				Класифікація об'єктів у випадку невідомих розподілень даних	3.05.04.05
				Методи оцінювання помилок класифікації	3.05.04.06

Шифр навчальної дисципліни	Назва навчальної дисципліни	Назва блока змістового модуля	Шифр блока змістових модулів	Назва теми (змістового модуля)	Шифр теми (змістового модуля)
		Методи пошуку шаблонів даних	3.05.05	Асоціаційні правила. Послідовне відображення шаблонів даних	3.05.05.01
				Метод Apriori, побудова FP-дерев пошуку шаблонів даних.	3.05.05.02
				Min-max асоціації у базах даних	3.05.05.03
				Побудова hash-дерев	3.05.05.04
				Розробка OLAP-кубів під час аналізу багатомірних даних у великих БД.	3.05.05.05
				Способи та методи візуального відображення даних	3.05.05.06
		OLAP і Data Mining	3.05.06	Методи, стадії, задачі Data Mining	3.05.06.07
				Упровадження Data Mining, OLAP і сховищ даних у СППР	3.05.06.08
				Процес Data Mining	3.05.06.01
				Стандарти Data Mining	3.05.06.02
				Інструменти Data Mining	3.05.06.03
3.06	Веб-технології та веб-дизайн	Основи Веб	3.06.01	Структура і принципи Веб	3.06.01.01
				Уведення в клієнт-серверні технології Веб	3.06.01.02
				Протокол HTTP	3.06.01.03
				Клієнтські сценарії та застосування	3.06.01.04
				Серверні веб-застосування	3.06.01.05
		Веб-програмування	3.06.02	JavaScript. Програмна взаємодія з HTML документами на основі DOM API	3.06.02.01
				Мови розроблення сценаріїв Perl, PHP, JSP	3.06.02.02
				Розробка CGI-застосувань на Perl, PHP, JSP	3.06.02.03
				Основи розробки веб-застосувань з допомогою ASP.NET, J2EE	3.06.02.04
				Інтерфейси взаємодії веб-застосувань з СКБД	3.06.02.05
				Веб-сервіси та мови їх описування	3.06.02.06
		Основи XML	3.06.03	Мови описування схем XML	3.06.03.01
				DOM XML. Перетворення XML-документів	3.06.03.02
				Програмна обробка XML-документів з допомогою XML DOM	3.06.03.03
				Форматування і перетворення XML-документа з допомогою CSS та XSL. XSLT перетворення XML-документа	3.06.03.04
		Веб-портали	3.06.04	Інтеграція та взаємодія у веб-мережі	3.06.04.01
				Розробка веб-служби в ASP.NET, J2EE	3.06.04.02
				Розробка веб-контента. CMS/CMF	3.06.04.03
				Розробка RSS-джерел і RSS-рідерів	3.06.04.04
		Технологія AJAX	3.06.05	Уведення в технологію AJAX. Розробка мобільних веб-застосувань	3.06.05.01
				Реалізація асинхронної взаємодії браузера з веб-сервером з допомогою технології AJAX	3.06.05.02



1. Программа дисциплины

Змістовий модуль 1. Основи інтелектуального аналізу даних.
Тема 1. Методи первісної обробки даних.
Тема 2. Методи дослідження структури даних: візуалізація та автоматичне групування даних.
Змістовий модуль 2. Методи використання навчальної інформації.
Тема 1. Кореляційний і регресійний аналіз даних. Множинний регресійний аналіз.
Тема 2. Лінійна множинна регресійна модель. Перевірка адекватності моделі.
Тема 3. Нелінійне оцінювання параметрів.
Змістовий модуль 3. Методи багатомірного розвідувального аналізу.
Тема 1. Кластерний аналіз. Ієрархічна та секційна кластеризації.
Тема 2. Методи кластеризації: процедура Мак-Кина, метод k-means, сітчасті методи.
Тема 3. Растрова кластеризація об’єктів.
Тема 4. Лінійний дискримінантний аналіз. Побудова канонічних та класифікаційних функцій.
Змістовий модуль 4. Методи класифікації та прогнозування.
Тема 1. Дерева рішень.
Тема 2. Методи опорних векторів, «найближчого сусіда», Байеса.
Тема 3. Аналіз багатомірних угруповань.
Тема 4. Статистична обробка тимчасових рядів і прогнозування.
Тема 5. Класифікація об’єктів у випадку невідомих розподілень даних.
Тема 6. Методи оцінювання помилок класифікації.

1. Программа дисциплины

Змістовий модуль 5. Методи пошуку шаблонів даних.
Тема 1. Асоціаційні правила. Послідовне відображення шаблонів даних.
Тема 2. Метод Apriori, побудова FP-дерев пошуку шаблонів даних.
Тема 3. Min-max асоціації у базах даних.
Тема 4. Побудова hash-дерев.
Тема 5. Розробка OLAP-кубів під час аналізу багатомірних даних у великих БД.
Тема 6. Способи та методи візуального відображення даних.
Змістовий модуль 6. OLAP і Data Mining.
Тема 1. Методи, стадії, задачі Data Mining.
Тема 2. Упровадження Data Mining, OLAP і сховищ даних у СППР.
Тема 3. Процес Data Mining.
Тема 4. Стандарти Data Mining.
Тема 5. Інструменти Data Mining.

1. Программа дисциплины

Барсегян А. А. Анализ данных и процессов : учеб. пособие / А. А. Барсегян, А. А. Куприянов [и др.]. - 3-е изд. - СПб. : БХВ - Петербург, 2009. - 512 с.

Статистика / В. С. Мхитарян [и др.] ; ред. В. С. Мхитарян. - 11-е изд., стер. - М. : ACADEMIA, 2012. - 272 с.

Алешин Л.И. Методы аналитической обработки данных / Л. И. Алешин, Ю. С. Гузев. - М. : Литера, 2008. - 144 с.

Чубукова И.А. Data Mining : учеб. пособие / И. А. Чубукова. - 2-е изд., испр. - М. : ИНТУИТ : БИНОМ. ЛЗ, 2008. - 382 с.

Халафян А. А. Статистический анализ данных / А. А. Халафян. - Изд. 2-е. - М. : БИНОМ, 2010. - 528 с.

Руденко О. Г. Искусственные нейронные сети / О. Г. Руденко, Е. В. Бодянский. - Харьков : Компания СМИТ, 2005. - 408 с.

Куликов Е. И. Прикладной статистический анализ / Е. И. Куликов. - 2-е изд., перераб. и доп. - М. : Горячая линия - Телеком, 2008. - 464 с.

Допоміжна література

Елисеева, И. И. Общая теория статистики / И. И. Елисеева, М. М. Юзбашев ; ред. И. И. Елисеева. - 4-е изд., перераб. и доп. - М. : Финансы и статистика, 2003. - 480 с.

Мандель, И. Д. Кластерный анализ / И. Д. Мандель. - М. : Финансы и статистика, 1988. - 176 с.

Статистические методы для ЭВМ / Ред. К. Энслейн, Э. Рэлстон, Г. С. Уилф. - М. : Наука, 1986. - 460 с.

Енюков, И. С. Методы, алгоритмы, программы многомерного статистического анализа / И. С. Енюков. - М. : Финансы и статистика, 1986. - 232 с.

1. Методы первичной обработки данных

1. Проблемы количественных измерений

2. Этапы и процедуры первичной обработки данных

3. Методы математической и статистической обработки данных

4. Проверка гипотез

5. Табличное и графическое представление данных

Способы статистического наблюдения

- Непосредственное наблюдение
- Документальное наблюдение или документальная запись
- Опрос

1. Методы первичной обработки данных

1. **История развития статистики как науки.**
2. **Основные понятия и методы статистики.**
3. **Структура отраслей статистической науки.***
4. **Организация государственной системы статистики РФ.***

1. Методы первичной обработки данных

	Атрибуты				
	Код клиента	Возраст	Семейное положение	Доход	Класс
Объекты	1	18	Single	125	1
	2	22	Married	100	1
	3	30	Single	70	1
	4	32	Married	120	1
	5	24	Divorced	95	2
	6	25	Married	60	1
	7	32	Divorced	220	1
	8	19	Single	85	2
	9	22	Married	75	1
	10	40	Single	90	2

1. Методы первичной обработки данных

Измерение - процесс присвоения чисел характеристикам изучаемых объектов согласно определенному правилу.

В процессе подготовки данных измеряется не сам объект, а его характеристики.

Шкала - правило, в соответствии с которым объектам присваиваются числа.

1. Методы первичной обработки данных

Измерение - процесс присвоения чисел характеристикам изучаемых объектов согласно определенному правилу.

В процессе подготовки данных измеряется не сам объект, а его характеристики.

Шкала - правило, в соответствии с которым объектам присваиваются числа.

1. Методы первичной обработки данных

	Атрибуты				
	Код клиента	Возраст	Семейное положение	Доход	Класс
Объекты	1	18	Single	125	1
	2	22	Married	100	1
	3	30	Single	70	1
	4	32	Married	120	1
	5	24	Divorced	95	2
	6	25	Married	60	1
	7	32	Divorced	220	1
	8	19	Single	85	2
	9	22	Married	75	1
	10	40	Single	90	2

1. Методы первичной обработки данных

Измерение - процесс присвоения чисел характеристикам изучаемых объектов согласно определенному правилу.

В процессе подготовки данных измеряется не сам объект, а его характеристики.

Шкала - правило, в соответствии с которым объектам присваиваются числа.

1. Методы первичной обработки данных

	Атрибуты				
	Код клиента	Возраст	Семейное положение	Доход	Класс
Объекты	1	18	Single	125	1
	2	22	Married	100	1
	3	30	Single	70	1
	4	32	Married	120	1
	5	24	Divorced	95	2
	6	25	Married	60	1
	7	32	Divorced	220	1
	8	19	Single	85	2
	9	22	Married	75	1
	10	40	Single	90	2

1. Методы первичной обработки данных

Порядковая шкала (ordinal scale) - шкала, в которой числа присваивают объектам для обозначения относительной позиции объектов, но не величины различий между ними.

Интервальная шкала (interval scale) - шкала, разности между значениями которой могут быть вычислены, однако их отношения не имеют смысла.

Интервальная шкала (interval scale) - шкала, разности между значениями которой могут быть вычислены, однако их отношения не имеют смысла.

Дихотомическая шкала (dichotomous scale) - шкала, содержащая только две категории.

Методы исследования структуры данных: визуализация и автоматическая группировка данных

Data Mining - это процесс поддержки принятия решений, основанный на поиске в данных скрытых закономерностей (шаблонов информации) [3].

Технологию Data Mining достаточно точно определяет Григорий Пиатецкий-Шапиро (Gregory Piatetsky-Shapiro) - один из основателей этого направления:

Data Mining - это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Суть и цель технологии Data Mining можно охарактеризовать так: это технология, которая предназначена для поиска в больших объемах данных неочевидных, объективных и полезных на практике закономерностей.

1. Методы первичной обработки данных

Неочевидных - это значит, что найденные закономерности не обнаруживаются стандартными методами обработки информации или экспертным путем.

Объективных - это значит, что обнаруженные закономерности будут полностью соответствовать действительности, в отличие от экспертного мнения, которое всегда является субъективным.

Практически полезных - это значит, что выводы имеют конкретное значение, которому можно найти практическое применение.

2. Корреляционный и регрессионный анализ данных. Множественный регрессионный анализ.

- **Методы выявления корреляционной зависимости**

- Построение и анализ параллельных рядов. При этом строится ранжированный ряд значений факторного признака и параллельно – ряд соответствующих значений признака-результата. По согласованному или несогласованному изменению значений фактора и результата судят о наличии либо отсутствии зависимости.

- Построение и анализ групповых таблиц. Групповая таблица строится по правилам аналитической группировки. В качестве группировочного признака используется факторный признак. По каждой из выделенных групп рассчитывается среднее значение результативного признака. Наличие закономерности в изменении средних величин зависимой переменной будет свидетельствовать о присутствии корреляционной связи.

2. Корреляционный и регрессионный анализ данных. Множественный регрессионный анализ.

- Построение и анализ корреляционных таблиц. В отличие от групповых, построение корреляционных таблиц предполагает группировку данных и по признаку-фактору, и по признаку-результату. На пересечении строк и столбцов проставляют частоты, т.е. число единиц совокупности с данным сочетанием уровней изучаемых признаков. Характер расположения частот на поле таблицы позволяет выдвинуть предположение о наличии и направлении зависимости между признаками.

- Графический метод. В прямоугольной системе координат по оси абсцисс откладываются значения признака-фактора, а по оси ординат – значения результативного признака. Точки на графике соответствуют единицам совокупности с конкретными сочетаниями значений признаков. Получаемый точечный график называют "полем корреляции". По расположению точек на графике судят о наличии или отсутствии зависимости, а также о направлении и степени тесноты корреляционной связи.

Корреляционный анализ

- Первой и простейшей характеристикой тесноты связи является линейный коэффициент парной корреляции.
- Показатели корреляции основаны на оценке сопряженной вариации изучаемых признаков. Парный коэффициент корреляции *(r)* – это нормированный коэффициент ковариации. Ковариация, являясь мерой взаимосвязи двух переменных, рассчитывается как средняя величина произведения отклонений индивидуальных значений анализируемых признаков от их средних значений:

$$Cov(y,x)=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x}) \qquad (1)$$

2. Корреляционный и регрессионный анализ данных. Множественный регрессионный анализ.

Парный коэффициент корреляции:

$$r = \frac{\sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right)}{n} \qquad (2)$$

где n – число единиц в статистической совокупности, σy – среднее квадратическое отклонение признака-результата; σx – среднее квадратическое отклонение признака-фактора.

Линейный коэффициент корреляции Пирсона:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n\sigma_x\sigma_y} \qquad (3)$$

или

$$r = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x\sigma_y} \qquad (4)$$

Коэффициент корреляции изменяется в пределах

$$0 \le |r| \le 1 \qquad (5)$$

Если $r = 0$, линейная связь между изучаемыми признаками отсутствует. Если $|r| = 1$, связь функциональная, т.е. значение зависимой переменной полностью определяется независимой переменной. Положительное значение коэффициента свидетельствует о прямой зависимости между признаками, отрицательная – об обратной.

2. Корреляционный и регрессионный анализ данных. Множественный регрессионный анализ.

Парный коэффициент корреляции – это симметричная характеристика, т.е. $r_{yx} = r_{xy}$.

Значение r отражает только степень тесноты корреляционной связи между изучаемыми признаками, но не свидетельствует о причинно-следственной зависимости между ними.

Обоснование наличия причинно-следственной связи между признаками опирается на анализ природы изучаемого явления.

2. Корреляционный и регрессионный анализ данных. Множественный регрессионный анализ.

- Квадрат коэффициента корреляции (r^2) называется коэффициентом детерминации. Его значение изменяется в пределах от 0 до 1, и означает долю вариации результативного признака, обусловленную вариацией признака-фактора.

- Парный коэффициент корреляции достаточно точно оценивает тесноту связи в условиях **линейной зависимости** между изучаемыми признаками. При наличии **нелинейной связи** он может привести к неверным выводам о степени тесноты связи (его значение занижено) .

Парный регрессионный анализ

- Сутью регрессионного анализа является описание "технологии" влияния признаков-факторов на признак-результат, который в конкретных практических задачах выступает объектом управления.
- Регрессионный анализ предполагает теоретический анализ природы изучаемого явления с целью определения круга факторов, оказывающих влияние на поведение результативного признака. На базе корреляционного анализа выявляется наличие статистически значимых связей в конкретных условиях места и времени. Затем строится уравнение регрессии (аналитическая форма изучаемой зависимости), которое при определенных условиях может быть признано статистической моделью связи между признаками.

2. Корреляционный и регрессионный анализ данных. Множественный регрессионный анализ.

- Уравнение регрессии – это математическая функция, описывающая зависимость условного среднего значения результативной (зависимой) переменной от заданных значений факторных (независимых) переменных. Таким образом, уравнение регрессии отражает основную тенденцию связи, характерную для изучаемой статистической совокупности в целом.

2. Корреляционный и регрессионный анализ данных. Множественный регрессионный анализ.

В регрессионном анализе можно выделить три составляющие:

- - определение типа функции (структуры модели) для описания изучаемой зависимости;
- - расчет неизвестных параметров уравнения регрессии;
- - оценку качества модели.

До широкого распространения компьютерных технологий перечисленные элементы являлись последовательными этапами анализа. В современных условиях все процедуры выполняются комплексно.

2. Корреляционный и регрессионный анализ данных. Множественный регрессионный анализ.

Первый этап регрессионного анализа – поиск линии регрессии, которая бы лучшим образом аппроксимировала поле корреляции.

Необходимо учитывать природу изучаемых показателей, специфику их взаимосвязи, свойства математических функций.

Современные ППП позволяют одновременно построить несколько видов уравнений, а затем, пользуясь специальными критериями, отобрать лучшую модель.

В качестве критерия могут быть использованы:

максимальное значение коэффициента детерминации, максимальное значение *F*-критерия Фишера, минимальное значение остаточной дисперсии, минимальное значение стандартной ошибки уравнения, минимальное значение средней ошибки аппроксимации.

2. Корреляционный и регрессионный анализ данных. Множественный регрессионный анализ.

Для аналитического описания связи между признаками могут быть использованы следующие виды уравнений:

- 1) $\bar{y} = a_0 + a_1 x$ – прямая, линейная функция;
- 2) $\bar{y} = a_0 + a_1 x^2 + a_2 x$ – парабола;
- 3) $\bar{y} = a_0 + a_1 \frac{1}{x}$ – гипербола;
- 4) $\bar{y} = a_0 x^{a_1}$ – степенная функция;
- 5) $\bar{y} = \exp(a_0 + a_1 x)$ – экспонента и др.

2. Корреляционный и регрессионный анализ данных. Множественный регрессионный анализ.

Некоторые задачи корреляционно-регрессионного анализа, а также возможности ППП, делают необходимым выполнение операции линеаризации уравнений, т.е. приведение их к линейному виду путем логарифмирования. Производится замена признака-фактора и признака-результата их натуральными логарифмами. При проведении анализа с использованием линеаризации необходимо помнить о том, что все показатели и графические изображения рассчитываются и строятся для логарифмов признаков.

2. Корреляционный и регрессионный анализ данных. Множественный регрессионный анализ.

Простейшим видом уравнения регрессии является парная линейная регрессия

$$\bar{y} = a_0 + a_1 x + \varepsilon ,$$

где $\bar{y}$ – расчетное, теоретическое значение признака-результата;

a_0, a_1 – параметры уравнения регрессии;

ε – случайная величина.

Присутствие в уравнении ε связано с рядом причин, среди которых: наличие признаков-факторов, не включенных в данное уравнение; неправильное описание структуры модели; ошибки измерений и др.