Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования «Ростовский государственный экономический университет (РИНХ)»

УТВЕРЖДАЮ
Директор Института магистратуры
_____ Иванова Е.А.
«30» 08 2021 г.

Рабочая программа дисциплины
Методы анализа больших данных (Big Data)

Направление 38.04.01 Экономика
магистерская программа 38.04.01.09 "Финансовый аналитик"

Для набора _2021_ года

Квалификация
магистр

КАФЕДРА **Статистики, эконометрики и оценки рисков**

Распределение часов дисциплины по курсам

Курс	1		Итого	
Вид занятий	УП	РП		
Лекции	6	6	6	6
Лабораторные	6	6	6	6
Практические	6	6	6	6
Итого ауд.	18	18	18	18
Контактная работа	18	18	18	18
Сам. работа	50	50	50	50
Часы на контроль	4	4	4	4
Итого	72	72	72	72

ОСНОВАНИЕ

Учебный план утвержден учёным советом вуза от 30.08.2021 протокол № 1.

Программу составил(и): к.э.н., доцент, Кракашова О.А. _____

Зав. кафедрой: д.э.н., профессор, Ниворожкина Л.И. _____

Методическим советом направления: д.э.н., профессор, Ниворожкина Л.И. _____

1. ЦЕЛИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

1.1	Сформировать у студентов системное представление о технологиях многомерного анализа данных, интеллектуального анализа данных (Data Mining), их применении и инструментах, изучить основные методы прикладного анализа данных, развить навыки исследования различных процессов с использованием современных информационно-коммуникационных технологий, практического применения методов многомерного анализа и Data Mining для решения различных научных и технических задач в экономике и бизнесе.

2. ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ

УК-4:Способен применять современные коммуникативные технологии, в том числе на иностранном(ых) языке(ах), для академического и профессионального взаимодействия

ПК-4:Способен формировать информационную базу бизнес-анализа, оценивать текущее состояние организации (объекта исследования) выявлять, и оценивать несоответствия между параметрами ее текущего и будущего состояний

В результате освоения дисциплины обучающийся должен:

Знать:

современные методы и инструменты решения задач Data Mining и многомерного анализа больших данных;
базовые понятия технологии Big Data.

Уметь:

применять программные и аппаратные средства персонального компьютера для анализа больших данных;
анализировать большие массивы данных, характеризующие различные социально-экономические процессы.

Владеть:

методами обработки больших массивов информации (Big data) и анализа данных различной природы;
современными технологиями создания и анализа больших данных.

3. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

Код занятия	Наименование разделов и тем /вид занятия/	Семестр / Курс	Часов	Компетен-ции	Литература
	Раздел 1. Сбор, хранение и анализ больших данных				
1.1	Обзор Big Data. Методы и средства. Используемые программы. Технологии хранения больших данных. Определение источника больших данных. Исследование источника данных. Хранилище данных. /Лек/	1	2	УК-4 ПК-4	Л1.1 Л1.2 Л1.3 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8 Л1.9 Л1.10Л2.1 Л2.2 Л2.3 Л2.4 Л2.5 Л2.6 Л2.7 Л2.16 Л2.18 Л2.20 Л2.21 Л2.22 Л2.23 Л2.24Л3.1 Э1 Э2 Э3 Э4 Э5 Э6 Э7 Э8 Э9 Э10 Э11 Э12 Э13 Э14 Э15 Э16 Э17 Э18 Э19 Э20 Э21 Э22 Э23 Э24 Э25 Э26 Э27
1.2	Введение в R. R: начало работы, ввод данных и работа с большими массивами данных. Работа с электронными таблицами больших данных в Calc. /Лаб/	1	2	УК-4 ПК-4	Л1.1 Л1.2 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8 Л1.9 Л1.10Л2.1 Л2.2 Л2.4 Л2.5 Л2.6 Л2.7 Л2.16 Л2.18 Л2.20 Л2.21 Л2.22 Л2.23 Л2.24Л3.1 Э2 Э3 Э4 Э5 Э6 Э7 Э8 Э9 Э10 Э11 Э12 Э13 Э14 Э15 Э16 Э17 Э18 Э19 Э20 Э21 Э22 Э23 Э24 Э25 Э26 Э27

1.3	Методы и средства анализа Big Data. Используемые программы. Технологии хранения больших данных. Введение в R (ввод данных, работа с большими массивами данных). Определение источника больших данных. Исследование источника данных. Хранилище данных. /Ср/	1	4	УК-4 ПК-4	Л1.1 Л1.2 Л1.3 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8 Л1.9 Л1.10Л2.1 Л2.2 Л2.4 Л2.5 Л2.6 Л2.7 Л2.16 Л2.18 Л2.20 Л2.21 Л2.22 Л2.23 Л2.24Л3.1 Э1 Э2 Э3 Э4 Э5 Э6 Э7 Э8 Э9 Э10 Э11 Э12 Э13 Э14 Э15 Э16 Э17 Э18 Э19 Э20 Э21 Э22 Э24 Э25 Э26 Э27
1.4	R: визуализация Big Data, фиктивные переменные, прогнозы, проверка гипотез и ловушка дамми-переменных. Большие данные и скрипты. Проблемы регрессионного анализа больших данных (мультиколлинеарность и гетероскедастичность) в R. /Лаб/	1	2	УК-4 ПК-4	Л1.1 Л1.2 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8 Л1.9 Л1.10Л2.1 Л2.2 Л2.4 Л2.5 Л2.6 Л2.7 Л2.16 Л2.18 Л2.20 Л2.21 Л2.22 Л2.23 Л2.24Л3.1 Э1 Э2 Э3 Э4 Э5 Э6 Э7 Э8 Э9 Э10 Э11 Э12 Э13 Э14 Э15 Э16 Э17 Э18 Э19 Э20 Э21 Э22 Э24 Э25 Э26 Э27
1.5	Источник и хранилище больших данных. Большие данные в R: постороение графиков и прогнозов, доверительных и предиктивных интервалов. /Ср/	1	4	УК-4 ПК-4	Л1.1 Л1.2 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8 Л1.9 Л1.10Л2.1 Л2.2 Л2.4 Л2.5 Л2.6 Л2.7 Л2.16 Л2.18 Л2.20 Л2.21 Л2.22 Л2.23 Л2.24Л3.1 Э1 Э2 Э3 Э4 Э5 Э6 Э7 Э8 Э9 Э10 Э11 Э12 Э13 Э14 Э15 Э16 Э17 Э18 Э19 Э20 Э21 Э22 Э24 Э25 Э26 Э27
1.6	Процесс анализа больших данных. Технологии анализа больших данных. Научные проблемы в области больших данных. /Лек/	1	2	УК-4 ПК-4	Л1.1 Л1.2 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8 Л1.9 Л1.10Л2.1 Л2.2 Л2.3 Л2.4 Л2.5 Л2.6 Л2.7 Л2.16 Л2.18 Л2.20 Л2.21 Л2.22 Л2.23 Л2.24Л3.1 Э1 Э2 Э3 Э4 Э5 Э6 Э7 Э8 Э9 Э10 Э11 Э12 Э13 Э14 Э15 Э16 Э17 Э18 Э19 Э20 Э21 Э22 Э24 Э25 Э26 Э27
1.7	Большие данные и скрипты. Проблемы регрессионного анализа больших данных (мультиколлинеарность и гетероскедастичность) в R. /Пр/	1	2	УК-4 ПК-4	Л1.1 Л1.2 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8 Л1.9 Л1.10Л2.1 Л2.2 Л2.4 Л2.5 Л2.6 Л2.7 Л2.16 Л2.18 Л2.20 Л2.21 Л2.22 Л2.23 Л2.24Л3.1 Э1 Э2 Э3 Э4 Э5 Э6 Э7 Э8 Э9 Э10 Э11 Э12 Э13 Э14 Э15 Э16 Э17 Э18 Э19 Э20 Э21 Э22 Э24 Э25 Э26 Э27

1.8	Процесс анализа больших данных. Технологии анализа больших данных. Научные проблемы в области больших данных. /Ср/	1	4	УК-4 ПК-4	Л1.1 Л1.2 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8 Л1.9 Л1.10Л2.1 Л2.2 Л2.4 Л2.5 Л2.6 Л2.7 Л2.16 Л2.18 Л2.20 Л2.21 Л2.22 Л2.23 Л2.24Л3.1 Э1 Э2 Э3 Э4 Э5 Э6 Э7 Э8 Э9 Э10 Э11 Э12 Э13 Э14 Э15 Э16 Э17 Э18 Э19 Э20 Э21 Э22 Э24 Э25 Э26 Э27
	Раздел 2. Методы и модели анализа больших данных				
2.1	Прогнозирование и предвидение в социально-политических и медиа процессах. Методы прогнозирования, использующие большие данные. /Лек/	1	2	УК-4 ПК-4	Л1.1 Л1.2 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8 Л1.9 Л1.10Л2.1 Л2.2 Л2.3 Л2.4 Л2.5 Л2.6 Л2.7 Л2.16 Л2.18 Л2.20 Л2.21 Л2.22 Л2.23 Л2.24Л3.1 Э1 Э2 Э3 Э4 Э5 Э6 Э7 Э8 Э9 Э10 Э11 Э12 Э13 Э14 Э15 Э16 Э17 Э18 Э19 Э20 Э21 Э22 Э24 Э25 Э26 Э27
2.2	Прогнозное моделирование: работа с регрессионными моделями больших данных в Calc и R. /Пр/	1	2	УК-4 ПК-4	Л1.1 Л1.2 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8 Л1.9 Л1.10Л2.1 Л2.2 Л2.4 Л2.5 Л2.6 Л2.7 Л2.16 Л2.18 Л2.20 Л2.21 Л2.22 Л2.23 Л2.24Л3.1 Э1 Э2 Э3 Э4 Э5 Э6 Э7 Э8 Э9 Э10 Э11 Э12 Э13 Э14 Э15 Э16 Э17 Э18 Э19 Э20 Э21 Э22 Э24 Э25 Э26 Э27
2.3	Методы прогнозирования в социально-экономических процессах. Модели прогнозирования: нейронные сети. /Ср/	1	10	УК-4 ПК-4	Л1.1 Л1.2 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8 Л1.9 Л1.10Л2.1 Л2.2 Л2.3 Л2.4 Л2.5 Л2.6 Л2.7 Л2.16 Л2.18 Л2.20 Л2.21 Л2.22 Л2.23 Л2.24Л3.1 Э1 Э2 Э3 Э4 Э5 Э6 Э7 Э8 Э9 Э10 Э11 Э12 Э13 Э14 Э15 Э16 Э17 Э18 Э19 Э20 Э21 Э22 Э24 Э25 Э26 Э27
2.4	R: даты и временные ряды, загрузка больших данных и тесты на автокорреляцию, качественные переменные, предельные эффекты и ROC кривая. Экспоненциальное сглаживание Холта с корректировкой тренда в моделях временных рядов больших данных в Calc и R. Примеры анализа временных рядов больших данных в R.Определение выбросов. /Лаб/	1	2	УК-4 ПК-4	Л1.1 Л1.2 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8 Л1.9 Л1.10Л2.1 Л2.2 Л2.4 Л2.5 Л2.6 Л2.7 Л2.16 Л2.18 Л2.20 Л2.21 Л2.22 Л2.23 Л2.24Л3.1 Э1 Э2 Э3 Э4 Э5 Э6 Э7 Э8 Э9 Э10 Э11 Э12 Э13 Э14 Э15 Э16 Э17 Э18 Э19 Э20 Э21 Э22 Э24 Э25 Э26 Э27

2.5	OLAP-системы. /Ср/	1	2	УК-4 ПК-4	Л1.1 Л1.5 Л1.6 Л1.7 Л1.8 Л1.9 Л1.10Л2.16 Л2.23 Л2.24 Э1 Э2 Э3 Э4 Э5 Э6 Э7 Э8 Э9 Э10 Э11 Э12 Э13 Э14 Э15 Э16 Э17 Э18 Э19 Э20 Э21 Э22 Э24 Э25 Э26 Э27
2.6	Интеллектуальный анализ данных (Data Mining). /Ср/	1	4	УК-4 ПК-4	Л1.1 Л1.2 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8 Л1.9 Л1.10Л2.1 Л2.2 Л2.3 Л2.4 Л2.5 Л2.6 Л2.7 Л2.16 Л2.17 Л2.18 Л2.20 Л2.21 Л2.22 Л2.23 Л2.24Л3.1 Э1 Э2 Э3 Э4 Э5 Э6 Э7 Э8 Э9 Э10 Э11 Э12 Э13 Э14 Э15 Э16 Э17 Э18 Э19 Э20 Э21 Э22 Э24 Э25 Э26 Э27
2.7	Эконометрическое исследование на больших данных Российского мониторинга экономического положения и здоровья населения (РМЭЗ, он же RLMS). /Пр/	1	2	УК-4 ПК-4	Л1.1 Л1.2 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8 Л1.9 Л1.10Л2.1 Л2.2 Л2.4 Л2.5 Л2.6 Л2.7 Л2.16 Л2.18 Л2.19 Л2.20 Л2.21 Л2.22 Л2.23 Л2.24Л3.1 Э1 Э2 Э3 Э4 Э5 Э6 Э7 Э8 Э9 Э10 Э11 Э12 Э13 Э14 Э15 Э16 Э17 Э18 Э19 Э20 Э21 Э22 Э24 Э25 Э26 Э27
2.8	Эконометрическое исследование на больших данных Российского мониторинга экономического положения и здоровья населения (РМЭЗ, он же RLMS). /Ср/	1	12	УК-4 ПК-4	Л1.1 Л1.2 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8 Л1.9 Л1.10Л2.1 Л2.2 Л2.4 Л2.5 Л2.6 Л2.7 Л2.16 Л2.18 Л2.19 Л2.20 Л2.21 Л2.22 Л2.23 Л2.24Л3.1 Э1 Э2 Э3 Э4 Э5 Э6 Э7 Э8 Э9 Э10 Э11 Э12 Э13 Э14 Э15 Э16 Э17 Э18 Э19 Э20 Э21 Э22 Э24 Э25 Э26 Э27
2.9	Задачи и методы интеллектуального анализа больших данных. /Ср/	1	2	УК-4 ПК-4	Л1.1 Л1.2 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8 Л1.9 Л1.10Л2.1 Л2.2 Л2.3 Л2.4 Л2.5 Л2.6 Л2.7 Л2.16 Л2.18 Л2.20 Л2.21 Л2.22 Л2.23 Л2.24Л3.1 Э1 Э2 Э3 Э4 Э5 Э6 Э7 Э8 Э9 Э10 Э11 Э12 Э13 Э14 Э15 Э16 Э17 Э18 Э19 Э20 Э21 Э22 Э24 Э25 Э26 Э27

2.10	Кластерный анализ на больших данных. Анализ потребительской корзины. Использование метода k-средних для сегментирования клиентской базы. Сетевые графы и определение сообществ. /Ср/	1	2	УК-4 ПК-4	Л1.1 Л1.2 Л1.5 Л1.6 Л1.7 Л1.8 Л1.9 Л1.10Л2.1 Л2.2 Л2.4 Л2.5 Л2.6 Л2.7 Л2.16 Л2.18 Л2.20 Л2.21 Л2.22 Л2.23 Л2.24Л3.1 Э1 Э2 Э3 Э4 Э5 Э6 Э7 Э8 Э9 Э10 Э11 Э12 Э13 Э14 Э15 Э16 Э17 Э18 Э19 Э20 Э21 Э22 Э24 Э25 Э26 Э27
2.11	Задачи и методы интеллектуального анализа данных. Кластерный анализ на больших данных. /Ср/	1	2	УК-4 ПК-4	Л1.1 Л1.2 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8 Л1.9 Л1.10Л2.1 Л2.2 Л2.4 Л2.5 Л2.6 Л2.7 Л2.16 Л2.18 Л2.20 Л2.21 Л2.22 Л2.23 Л2.24Л3.1 Э1 Э2 Э3 Э4 Э5 Э6 Э7 Э8 Э9 Э10 Э11 Э12 Э13 Э14 Э15 Э16 Э17 Э18 Э19 Э20 Э21 Э22 Э24 Э25 Э26 Э27
2.12	Инструменты Data Mining. /Ср/	1	4	УК-4 ПК-4	Л1.1 Л1.2 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8 Л1.9 Л1.10Л2.1 Л2.2 Л2.3 Л2.4 Л2.5 Л2.6 Л2.7 Л2.16 Л2.18 Л2.20 Л2.21 Л2.22 Л2.23 Л2.24Л3.1 Э1 Э2 Э3 Э4 Э5 Э6 Э7 Э8 Э9 Э10 Э11 Э12 Э13 Э14 Э15 Э16 Э17 Э18 Э19 Э20 Э21 Э22 Э24 Э25 Э26 Э27
2.13	/Зачёт/	1	4	УК-4 ПК-4	Л1.1 Л1.2 Л1.3 Л1.4 Л1.5 Л1.6 Л1.7 Л1.8 Л1.9 Л1.10Л2.1 Л2.2 Л2.3 Л2.4 Л2.5 Л2.6 Л2.7 Л2.8 Л2.9 Л2.10 Л2.11 Л2.12 Л2.13 Л2.14 Л2.15 Л2.16 Л2.17 Л2.18 Л2.19 Л2.20 Л2.21 Л2.22 Л2.23 Л2.24Л3.1 Э1 Э2 Э3 Э4 Э5 Э6 Э7 Э8 Э9 Э10 Э11 Э12 Э13 Э14 Э15 Э16 Э17 Э18 Э19 Э20 Э21 Э22 Э23 Э24 Э25 Э26 Э27

4. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

Структура и содержание фонда оценочных средств для проведения текущей и промежуточной аттестации представлены в Приложении 1 к рабочей программе дисциплины.

5. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

5.1. Основная литература

	Авторы, составители	Заглавие	Издательство, год	Колич-во

	Авторы, составители	Заглавие	Издательство, год	Колич-во
Л1.1	Ратникова Т. А., Фурманов К. К.	Анализ панельных данных и данных о длительности состояний: учеб. пособие	М.: Издат. дом Высш. шк. экономики, 2014	20
Л1.2	Ниворожкина Л. И.	Статистические методы анализа данных: учеб.	М.: РИО, 2016	105
Л1.3	Ниворожкина Л. И.	Эконометрика: учеб. пособие на англ. яз.	Ростов н/Д: Изд-во РГЭУ (РИНХ), 2017	44
Л1.4	Герасимов А. Н., Громов Е. И., Скрипниченко Ю. С.	Эконометрика: продвинутый уровень: учебное пособие	Ставрополь: Ставропольский государственный аграрный университет (СтГАУ), 2016	https://biblioclub.ru/index.php?page=book&id=484978 неограниченный доступ для зарегистрированных пользователей
Л1.5	Неделько, В. М.	Основы статистических методов машинного обучения: учебное пособие	Новосибирск: Новосибирский государственный технический университет, 2010	http://www.iprbookshop.ru/45418.html неограниченный доступ для зарегистрированных пользователей
Л1.6	Лемешко, Б. Ю., Лемешко, С. Б., Постовалов, С. Н., Чимитова, Е. В.	Статистический анализ данных, моделирование и исследование вероятностных закономерностей. Компьютерный подход: монография	Новосибирск: Новосибирский государственный технический университет, 2011	http://www.iprbookshop.ru/47719.html неограниченный доступ для зарегистрированных пользователей
Л1.7	Гончарова, Н. Д., Терехова, Ю. С.	Анализ и моделирование статистических рядов: учебное пособие	Новосибирск: Сибирский государственный университет телекоммуникаций и информатики, 2016	http://www.iprbookshop.ru/69536.html неограниченный доступ для зарегистрированных пользователей
Л1.8	Дубина, И. Н.	Математико-статистические методы и инструменты в эмпирических социально-экономических исследованиях: учебное пособие	Саратов: Вузовское образование, 2018	http://www.iprbookshop.ru/76234.html неограниченный доступ для зарегистрированных пользователей
Л1.9	Шнарева, Г. В., Пономарева, Ж. Г.	Анализ данных: учебно-методическое пособие	Симферополь: Университет экономики и управления, 2019	http://www.iprbookshop.ru/89482.html неограниченный доступ для зарегистрированных пользователей
Л1.10	Брусенцев, А. Г.	Анализ данных и процессов. Ч.1. Методы статистического анализа данных: учебное пособие	Белгород: Белгородский государственный технологический университет им. В.Г. Шухова, ЭБС АСВ, 2017	http://www.iprbookshop.ru/92237.html неограниченный доступ для зарегистрированных пользователей

5.2. Дополнительная литература

	Авторы, составители	Заглавие	Издательство, год	Колич-во
Л2.1	Пересецкий А. А.	Эконометрические методы в дистанционном анализе деятельности российских банков	М.: Издат. дом Высш. шк. экономики, 2012	20
Л2.2	Арженовский С. В.	Эконометрика финансовых рынков: метод. указания по изучению дисциплины	Ростов н/Д: Изд-во РГЭУ (РИНХ), 2015	95
Л2.3	Ниворожкина Л. И.	Статистические методы в управлении рисками: анализ данных о длительности состояний: учеб.-метод. пособие	Ростов н/Д: Изд-во РГЭУ (РИНХ), 2015	268
Л2.4	Елисеева И. И.	Эконометрика: учеб. для бакалавриата и магистратуры	М.: Юрайт, 2016	60
Л2.5	Крутиков В.Н., Мешечкин В.В.	Анализ данных	Кемерово: КГУ, 2014	http://biblioclub.ru/index.php?page=book_red&id=278426&sr=1 неограниченный доступ для зарегистрированных пользователей

	Авторы, составители	Заглавие	Издательство, год	Колич-во
Л2.6	Айвазян С. А., Иванова С. С.	Эконометрика: учеб. пособие для вузов	М.: Маркет ДС, 2007	100
Л2.7	Герасимов А. Н., Громов Е. И., Скрипниченко Ю. С.	Эконометрика: учеб. пособие для студентов высш. учеб. заведений, обучающихся по напр. подгот. 38.03.01 "Экономика"	Ростов н/Д: Феникс, 2017	20
Л2.8		Хранилища данных. Лекция 1. Понятия о хранилищах. Презентация	Москва: Национальный Открытый Университет «ИНТУИТ», 2014	http://biblioclub.ru/index.php?page=book&id=237105 неограниченный доступ для зарегистрированных пользователей
Л2.9		Хранилища данных. Лекция 3. Создание куба в SQL Server 2005. Презентация	Москва: Национальный Открытый Университет «ИНТУИТ», 2014	http://biblioclub.ru/index.php?page=book&id=237113 неограниченный доступ для зарегистрированных пользователей
Л2.10		Хранилища данных. Лекция 4. Создание многомерного хранилища данных на основе MS SQL Server 2005. Презентация	Москва: Национальный Открытый Университет «ИНТУИТ», 2014	http://biblioclub.ru/index.php?page=book&id=237114 неограниченный доступ для зарегистрированных пользователей
Л2.11		Хранилища данных. Лекция 6. Работа с OLAP срезами. Презентация	Москва: Национальный Открытый Университет «ИНТУИТ», 2014	http://biblioclub.ru/index.php?page=book&id=237115 неограниченный доступ для зарегистрированных пользователей
Л2.12		Хранилища данных. Лекция 7. SQL Server – ProClarity. Презентация	Москва: Национальный Открытый Университет «ИНТУИТ», 2014	http://biblioclub.ru/index.php?page=book&id=237116 неограниченный доступ для зарегистрированных пользователей
Л2.13		Хранилища данных	Москва: Национальный Открытый Университет «ИНТУИТ», 2014	http://biblioclub.ru/index.php?page=book&id=237117 неограниченный доступ для зарегистрированных пользователей
Л2.14		Хранилища данных. Лекция 9. Обзор основных технологий и функциональных возможностей Crystal Analysis Professional 10.0. Презентация	Москва: Национальный Открытый Университет «ИНТУИТ», 2014	http://biblioclub.ru/index.php?page=book&id=237118 неограниченный доступ для зарегистрированных пользователей
Л2.15		Функциональное программирование. Лекция 11. Функциональные структуры данных	Москва: Национальный Открытый Университет «ИНТУИТ», 2014	http://biblioclub.ru/index.php?page=book&id=237268 неограниченный доступ для зарегистрированных пользователей
Л2.16	Алексеев В. Е., Таланов В. А.	Структуры данных. Модели вычислений: курс лекций	Москва: Национальный Открытый Университет «ИНТУИТ», 2016	https://biblioclub.ru/index.php?page=book&id=428782 неограниченный доступ для зарегистрированных пользователей

	Авторы, составители	Заглавие	Издательство, год	Колич-во
Л2.17	Нестеров С. А.	Интеллектуальный анализ данных средствами MS SQL Server 2008	Москва: Национальный Открытый Университет «ИНТУИТ», 2016	https://biblioclub.ru/index.php?page=book&id=429083 неограниченный доступ для зарегистрированных пользователей
Л2.18	Добронец Б. С., Попова О. А.	Численный вероятностный анализ неопределенных данных: монография	Красноярск: Сибирский федеральный университет (СФУ), 2014	https://biblioclub.ru/index.php?page=book&id=435672 неограниченный доступ для зарегистрированных пользователей
Л2.19	Рощина Я. М.	Основы моделирования экономического поведения домохозяйств на базе данных RLMS-HSE: лекции	Москва: Издательский дом Высшей школы экономики, 2015	http://biblioclub.ru/index.php?page=book&id=440284 неограниченный доступ для зарегистрированных пользователей
Л2.20	Крутиков В. Н., Мешечкин В. В.	Анализ данных: учебное пособие	Кемерово: Кемеровский государственный университет, 2014	https://biblioclub.ru/index.php?page=book&id=278426 неограниченный доступ для зарегистрированных пользователей
Л2.21		Прикладная эконометрика: журнал	Москва: Университет Синергия, 2017	https://biblioclub.ru/index.php?page=book&id=459346 неограниченный доступ для зарегистрированных пользователей
Л2.22		Прикладная эконометрика: журнал	Москва: Университет Синергия, 2018	https://biblioclub.ru/index.php?page=book&id=484968 неограниченный доступ для зарегистрированных пользователей
Л2.23	Маглеванный, И. И., Карякина, Т. И.	Математические основы первичной обработки экспериментальных данных: методические материалы по прикладной статистике	Волгоград: Волгоградский государственный социально-педагогический университет, «Перемена», 2015	http://www.iprbookshop.ru/40738.html неограниченный доступ для зарегистрированных пользователей
Л2.24	Гимазов, Р. М.	Статистическая обработка материалов исследования на компьютере: учебно-методическое пособие: направление подготовки 050100 педагогическое образование, профиль «физкультурное образование», «образование в области безопасности жизнедеятельности)»; направление подготовки 034400 физическая культура для лиц с отклонениями в состоянии здоровья (адаптивная физическая культура)	Сургут: Сургутский государственный педагогический университет, 2015	http://www.iprbookshop.ru/87033.html неограниченный доступ для зарегистрированных пользователей

5.3. Методические разрабоки

	Авторы, составители	Заглавие	Издательство, год	Колич-во
Л.1	Арженовский С. В., Торопова Т. В.	Эконометрическое моделирование с использованием пакетов прикладных программ: метод. указания к выполнению лаборатор. работ	Ростов н/Д: Изд-во РГЭУ (РИНХ), 2015	95

5.3 Профессиональные базы данных и информационные справочные системы

Для преподавания курса предполагается использование данных социологических опросов населения НАФИ, ВЦИОМ, ЦИРКОН, ФОМ, РОМИР, КОМКОН и других, а также данных репрезентативных опросов населения России, таких как, RLMS, NOBUS, GGS.

5.4. Перечень программного обеспечения

Libre Office, R.

5.5. Учебно-методические материалы для студентов с ограниченными возможностями здоровья

При необходимости по заявлению обучающегося с ограниченными возможностями здоровья учебно-методические материалы предоставляются в формах, адаптированных к ограничениям здоровья и восприятия информации. Для лиц с нарушениями зрения: в форме аудиофайла; в печатной форме увеличенным шрифтом. Для лиц с нарушениями слуха: в форме электронного документа; в печатной форме. Для лиц с нарушениями опорно-двигательного аппарата: в форме электронного документа; в печатной форме.

6. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

Помещения для проведения всех видов работ, предусмотренных учебным планом, укомплектованы необходимой специализированной учебной мебелью и техническими средствами обучения.

7. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ (МОДУЛЯ)

Методические указания по освоению дисциплины представлены в Приложении 2 к рабочей программе дисциплины.

Приложение 1

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

1. Описание показателей и критериев оценивания компетенций на различных этапах их формирования, описание шкал оценивания

1.1 Показатели и критерии оценивания компетенций:

ЗУН, составляющие компетенцию	Показатели оценивания	Критерии оценивания	Средства оценивания
ПК-1 Способность обобщать и критически оценивать результаты, полученные отечественными и зарубежными исследователями, выявлять перспективные направления, составлять программу исследований			
Знать современные методы и инструменты решения задач Data Mining и многомерного анализа больших данных.	Знает тенденции технологий интеллектуального анализа данных, стандартах и инструментах; типы закономерностей и сферы применения Data Mining; виды и способы организации хранилищ данных; классификацию аналитических систем; состав классов программных продуктов, образующих набор Business Intelligence. Аргументирует возможные ограничения применения методов, основные проблемы, возникающие при анализе данных, и пути их решения. Объясняет отличия Data Mining от классических статистических методов анализа и OLAP-систем.	Верно использует модели и методы эконометрики для решения задачи на больших данных с учетом ограничений	С, З, ЛР, Т
Уметь применять программные и аппаратные средства персонального компьютера для анализа больших данных	Умеет ориентироваться в современной системе источников информации. Видит и формулирует проблему. Самостоятельно ставит цели и задачи. Использует методы интеллектуального анализа данных и современные информационные технологии в своей профессиональной деятельности. Выбирает программное обеспечение согласно решаемой задаче.	Правильность применения команд в пакете прикладных программ, верные действия при моделировании и анализе больших данных	ЛР, З
Владеть методами обработки больших массивов информации (Big	Использует современную терминологию в области систем поддержки принятия решений и методологии решения задач в	Корректные методы для решения задачи, адекватная	З, ЛР

data) и анализа данных различной природы.	области многомерного анализа больших данных. Применяет современные программные пакеты многомерного анализа больших данных. Оценивает и формулирует выводы по результатам применения многомерного инструментария.	модель, верная интерпретация результатов моделирования	
ПК-9 Способность анализировать и использовать различные источники информации для проведения экономических расчетов			
Знать базовые понятия технологии Big Data.	Знает базовые понятия и технологии прогнозирования с использованием больших данных.	Верно использует модели и методы эконометрики для решения задачи на больших данных с учетом ограничений	С, З, ЛР, Т
Уметь анализировать большие массивы данных, характеризующие различные социально-эконом ические процессы.	Умеет определять массивы больших данных; анализировать кластеры больших данных. Строит различными способами прогнозы развития экономических процессов с использованием больших массивов данных.	Верно использует модели и методы эконометрики для решения задачи на больших данных с учетом ограничений	С, З, ЛР
Владеть современными технологиями создания и анализа больших данных.	Владеет терминологией курса, методологией и методикой прогнозирования.	Корректные методы для решения задачи, адекватная модель, верная интерпретация результатов моделирования	З, ЛР

1.2. Шкала оценивания:

Текущий контроль успеваемости и промежуточная аттестация осуществляется в рамках накопительной балльно-рейтинговой системы в 100-балльной шкале.

Промежуточная аттестация осуществляется по следующей шкале:

- 50-100 баллов (зачтено)
- 0-49 баллов (не зачтено).

2. Типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций в процессе освоения образовательной программы

Задания к зачету

ЗАДАНИЕ К ЗАЧЕТУ №1

1. Технологии Business Intelligence и реляционные системы управления базами данных.
2. Интегрированные процессы. Идентификация авторегрессионной модели скользящего среднего.

3. Задача.

Исходный файл с данными: EARNINGS.xls

В вашем распоряжении имеются следующие данные о 540 работниках (270 мужчин и 270 женщин): EARNINGS — текущий часовой заработок в долларах США, S — продолжительность обучения(число полных лет обучения), EXP — общий стаж работы после окончания учебы, FEMALE — пол респондента(0 — для мужчин, 1 — для женщин).

Импортируйте данные в R.

Постройте модель вида: $EARNINGS_i=\beta1+\beta2+\beta3EXP_i+\beta4FEMALE_i+\varepsilon_i$

Проведя тест Бреуша-Пагана, скажите, присутствует ли в модели гетероскедастичность, и если да, то с какой переменной она, скорее всего, связана (при ответе ориентируйтесь на значимость коэффициентов в соответствующем уравнении).

ЗАДАНИЕ К ЗАЧЕТУ № 2

1. Технологии обработки больших данных: NoSQL, MapReduce, Hadoop, R.
2. Коинтеграция. Анализ временных рядов.
3. Задача.

По 350 наблюдениям оценена модель зависимости заработной платы $wage_i$ (\$) от длительности обучения $schooling_i$ (годы) и опыта работы $experience_i$ (годы). Оцененная модель имеет вид: $\widehat{wage}_\iota = 400 + 25schooling_i + 60experience_i$. *ESS=130, TSS=210*. Исследователь решил добавить в модель образование родителей $mschooling_i$ и $fschooling_i$ (годы), после чего *ESS=180*. На уровне значимости 10% проверяя гипотезу о влиянии длительности обучения родителей на заработную плату их ребенка, укажите количество ограничений, которые приравнены к нулю в формулировке нулевой гипотезы?

ЗАДАНИЕ К ЗАЧЕТУ № 3

1. Понятие Большие данные. Роль цифровой информации в 21 веке.
2. Специальные методы анализа социально-политических и медиа процессов.
3. Задача.

По 400 наблюдениям оценена модель зависимости заработной платы $wage_i$ (\$) от длительности обучения $schooling_i$ (годы) и опыта работы $experience_i$ (годы). Оцененная модель имеет вид: $\widehat{wage}_\iota = 400 + 25schooling_i + 60experience_i$. ESS=125, TSS=200. Исследователь решил добавить в модель образование родителей $mschooling_i$ и $fschooling_i$ (годы), после чего ESS=175. На уровне значимости 1% проверяя гипотезу о влиянии длительности обучения родителей на заработную плату их ребенка, определите чему равно наблюдаемое значение тестовой статистики?

ЗАДАНИЕ К ЗАЧЕТУ № 4

1. Базовые принципы обработки больших данных.
2. Статистические оценки параметров. Доверительные области.
3. Задача.

Исследуется зависимость среднедушевого потребления алкоголя по странам мира от различных факторов.

Модель 1:

$ALCOi=\beta1+\beta2GDPi+\beta3MUSLi+\beta4BUDDi+\beta5HINDUi+\varepsilon i$,

где ALCOi — среднедушевое потребление чистого спирта на человека (л), GDPi — ВВП на душу населения (долларов США), MUSLi, BUDDi, HINDUi — доли населения исповедующего, соответственно, мусульманство, буддизм и индуизм (в % от общей численности населения). В ходе МНК-оценивания модели на основе данных о 50 странах получены следующие результаты: сумма квадратов остатков ESS=200, объясненная сумма квадратов RSS=300.

Также для проверки гипотезы о том, что религия не оказывает существенного влияния на потребление алкоголя, были оценены параметры второй модели:
Модель №2:
$ALCOi=\beta1+\beta2GDPi+\varepsilon i$.
Во второй модели, по сравнению с первой, значение RSS изменилось на100. Сколько составит скорректированный R^2 во второй модели?

ЗАДАНИЕ К ЗАЧЕТУ № 5

1. Предварительный анализ данных.
2. Основные возможности анализа больших данных в R.
3. Задача.
   Исходный файл с данными: EARNINGS.xls
   В вашем распоряжении имеются следующие данные о 540 работниках (270 мужчин и 270 женщин): EARNINGS — текущий часовой заработок в долларах США, S — продолжительность обучения(число полных лет обучения), EXP — общий стаж работы после окончания учебы, FEMALE — пол респондента(0 — для мужчин, 1 — для женщин).
   Импортируйте данные в R.
   Постройте модель вида: $EARNINGSi=\beta1+\beta2+\beta3EXPi+\beta4FEMALEi+\varepsilon i$.
   Оцените качество построенной модели и постройте доверительный и предиктивный интервалы для $\beta4$.

ЗАДАНИЕ К ЗАЧЕТУ № 6

1. Неметрические методы. Кластерный анализ. Дискриминантный анализ.
2. Коинтеграция. Анализ временных рядов.
3. Задача.
   Используем встроенный набор данных mtcars в RStudio. Сохраните в переменную логистическую регрессионную модель, где в качестве зависимой переменной выступает тип коробки передач (am), в качестве предикторов переменные disp, vs, mpg. Значения коэффициентов регрессии сохраните в переменную log_coef. Выпишите полученную спецификацию модели. Оцените ее качество.

ЗАДАНИЕ К ЗАЧЕТУ № 7

1. Основные возможности анализа больших данных в R.
2. Специальные методы анализа социально-политических и медиа процессов.
3. Задача.
   На встроенных в R данных npk, иллюстрирующими влияние применения различных удобрений на урожайность гороха (yield). Нашей задачей будет выяснить, существенно ли одновременное применение азота (фактор N) и фосфата (фактор P). Примените дисперсионный анализ, где будет проверяться влияние фактора применения азота (N), влияние фактора применения фосфата (P) и их взаимодействие. В ответе укажите p-value для взаимодействия факторов N и P.

ЗАДАНИЕ К ЗАЧЕТУ № 8

1. Теория моментов.
2. Многомерное шкалирование. Классическая модель многомерного шкалирования.
3. Задача.
   В переменной df сохранен subset данных mtcars только с переменными "wt", "mpg", "disp", "drat", "hp". Воспользуйтесь множественным регрессионным анализом, чтобы предсказать вес машины (переменная "wt"). Выберите такую комбинацию независимых переменных (из "mpg", "disp", "drat", "hp"), чтобы значение R^2 adjusted было наибольшим. Взаимодействия факторов учитывать не надо.

ЗАДАНИЕ К ЗАЧЕТУ № 9

1. Специальные методы анализа социально-политических и медиа процессов.

2. Дисперсионный анализ влияния качественных факторов. Ранговые методы.
3. Задача.

Воспользуйтесь встроенным датасетом attitude, чтобы предсказать рейтинг (rating) по переменным complaints и critical. Каково t-значение для взаимодействия двух факторов? Разделителем целой и дробной части в ответе должна быть запятая!

ЗАДАНИЕ К ЗАЧЕТУ № 10

1. Прогнозирование с моделями временных рядов. Доверительные интервалы прогноза.
2. Технологии обработки больших данных: NoSQL, MapReduce, Hadoop, R.
3. Задача.

На встроенном датасете LifeCycleSavings предсказать значение sr на основе всех остальных переменных в этом датасете. Напишите команду, которая создаёт линейную регрессию с главными эффектами и всеми возможными взаимодействиями второго уровня. Сохраните модель в переменную model. Выпишите итоговую спецификацию модели и оцените ее качество.

Критерии оценивания:

Максимальное количество баллов – 100.

Задание к зачету содержит 2 вопроса и 1 задачу, баллы и критерии оценивания по которым приведены выше. Баллы выставляются по каждому заданию в отдельности и суммируются.

Каждый теоретический вопрос оценивается отдельно, максимально в 24 балла.

Критерии оценивания отдельного вопроса:

- 13-24 балла. Ответ на вопрос верный; продемонстрировано наличие глубоких исчерпывающих / твердых и достаточно полных знаний, грамотное и логически стройное изложение материала при ответе, возможны отдельные погрешности и ошибки, уверенно исправленные и после дополнительных вопросов; продемонстрировано наличие глубоких исчерпывающих / твердых и достаточно полных знаний, грамотное и логически стройное изложение материала при ответе.
- 0-12 баллов. Ответ на вопрос лишь частично верен, продемонстрирована неточность и неуверенность ответов на дополнительные и наводящие вопросы, либо ответ на вопрос не верен, продемонстрирована неуверенность и неточность ответов на дополнительные и наводящие вопросы.

Задача оценивается максимально в 52 балла:

Критерии оценивания задачи:

- 27-52 балла. Задача решена в полном объеме, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов; либо задача решена в полном объеме с небольшими погрешностями, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов, в расчетах и выводах содержатся незначительные ошибки.
- 0-26 баллов. Задача решена частично, частично выбраны верные инструментальные методы и приемы решения, проведены частичные расчеты, сделан вывод по результатам проведенных расчетов с погрешностями либо задача не решена или решена частично, частично выбраны необходимые инструментальные методы и приемы решения, расчеты не проведены или проведены частично, вывод по результатам проведенных расчетов не сделан или ошибочен.

Зачет выставляется на основании итоговой суммы баллов, набранных студентом:

- 50-100 баллов «зачтено»;
- 0-49 баллов «не зачтено».

Тесты письменные и/или компьютерные
(Тестовые задания к экзамену и текущему контролю знаний)

Банк тестов

1. Банк тестов по модулям.

Модуль 1 «Сбор, хранение и анализ больших данных»

1.1 Задача классификации сводится к …

а) нахождения частых зависимостей между объектами или событиями;
б) определения класса объекта по его характеристикам;
в) определение по известным характеристикам объекта значение некоторого его параметра;
г) поиска независимых групп и их характеристик во всем множестве анализируемых данных.
1.2. Целью поиска ассоциативных правил является …
а) нахождения частых зависимостей между объектами или событиями;
б) определения класса объекта по его характеристикам;
в) определение по известным характеристикам объекта значение некоторого его параметра;
г) поиска независимых групп и их характеристик во всем множестве анализируемых данных.
1.3. Очистка данных — …
а) комплекс методов и процедур, направленных на устранение причин, мешающих корректной обработке: аномалий, пропусков, дубликатов, противоречий, шумов и т.д.;
б) процесс дополнения данных некоторой информацией, позволяющей повысить эффективность развязку аналитических задач;
в) объект, содержащий структурированные данные, которые могут оказаться полезными для решения аналитической задачи;
г) комплекс методов и процедур, направленных на извлечение данных из различных источников, обеспечение необходимого уровня их информативности и качества, преобразования в единый формат, в котором они могут быть загружены в хранилище данных или аналитическую систему.
1.4 Обогащение — …
а) комплекс методов и процедур, направленных на устранение причин, мешающих корректной обработке: аномалий, пропусков, дубликатов, противоречий, шумов и т.д. ;
б) процесс дополнения данных некоторой информацией, позволяющей повысить эффективность развязку аналитических задач;
в) объект, содержащий структурированные данные, которые могут оказаться полезными для решения аналитической задачи;
г) комплекс методов и процедур, направленных на извлечение данных из различных источников, обеспечение необходимого уровня их информативности и качества, преобразования в единый формат, в котором они могут быть загружены в хранилище данных или аналитическую систему.
1.5. Транзакция — …
а) некоторый набор операций над базой данных, который рассматривается как единственное завершенное, с точки зрения пользователя, действие над некоторой информацией, обычно связано с обращением к базе данных;
б) разновидность систем хранения, ориентирована на поддержку процесса анализа данных целостность, обеспечивает, непротиворечивость и хронологию данных, а также высокую скорость выполнения аналитических запросов;
в) высокоуровневые средства отражения информационной модели и описания структуры данных;
г) это установление зависимости дискретной выходной переменной от входных переменных.
1.6. Аналитическая платформа — …
а) специализированное программное решение (или набор решений), которое включает в себя все инструменты для извлечения закономерностей из сырых данных;
б) это группировка объектов (наблюдений, событий) на основе данных, описывающих свойства объектов;
в) набор данных, каждая запись которого представляет собой учебный пример, содержащий заданное входное влияние, что и отвечает ему правильный выходной результат;
г) подразделение искусственного интеллекта, изучающее методы построения алгоритмов, способных обучаться на данных.
1.7. Распределите нижеприведенные результаты внедрения Big Data по степени эффективности (наиболее эффективные – 1 место, наименее – 7).

1. Улучшение клиентского сервиса	
2. Улучшение реагирования на запросы клиентов	
3. Рост эффективности обработки клиентских запросов	
4. Улучшение интеграции в цепи поставок	
5. Оптимизация запасов и продуктивности основных активов	
6. Улучшение процессов планирования компании	

7. Оптимизация затрат на сервис	

1.8. Консолидация — …

а) комплекс методов и процедур, направленных на устранение причин, мешающих корректной обработке: аномалий, пропусков, дубликатов, противоречий, шумов и т.д.;

б) процесс дополнения данных некоторой информацией, позволяющей повысить эффективность развязку аналитических задач;

в) объект, содержащий структурированные данные, которые могут оказаться бесполезными для решения аналитической задачи;

г) комплекс методов и процедур, направленных на извлечение данных из различных источников, обеспечение необходимого уровня их информативности и качества, преобразования в единый формат, в котором они могут быть загружены в хранилище данных или аналитическую систему.

1.9. Виды физической неопределенности данных:

а) неточность измерений значений определенной величины, выполняемых физическими приборами; случайность (или наличие во внешней среде нескольких возможностей, каждая из которых случайным образом может стать действительностью);

б) неопределенность значений слов (многозначность, размытость, непонятность, нечеткость); неоднозначность смысла фраз (синтаксическая и семантическая);

в) случайность (или наличие во внешней среде нескольких возможностей, каждая из которых случайным образом может стать действительностью); неопределенность значений слов (многозначность, размытость, неясность, нечеткость);

г) неоднозначность смысла фраз (синтаксическая и семантическая).

1.10. Метаданные — …

а) некоторый набор операций над базой данных, который рассматривается как единственное завершенное, с точки зрения пользователя, действие над некоторой информацией, обычно связано с обращением к базе данных;

б) разновидность систем хранения, ориентирована на поддержку процесса анализа данных целостность, обеспечивает, непротиворечивость и хронологию данных, а также высокую скорость выполнения аналитических запросов;

в) высокоуровневые средства отражения информационной модели и описания структуры данных;

г) это установление зависимости дискретной выходной переменной от входных переменных.

1.11. Классификация — …

а) некоторый набор операций над базой данных, который рассматривается как единственное завершенное, с точки зрения пользователя, действие над некоторой информацией, обычно связано с обращением к базе данных;

б) разновидность систем хранения, ориентированная на поддержку процесса анализа данных, обеспечивает целостность, непротиворечивость и хронологию данных, а также высокую скорость выполнения аналитических запросов;

в) высокоуровневые средства отражения информационной модели и описания структуры данных;

г) это установление зависимости дискретной выходной переменной от входных переменных.

1.12. Регрессия — …

а) это установление зависимости непрерывной выходной переменной от входных переменных;

б) эта группировка объектов (наблюдений, событий) на основе данных, описывающих свойства объектов;

в) выявление закономерностей между связанными событиями;

г) это установление зависимости дискретной выходной переменной от входных переменных.

1.13. Кластеризация — …

а) это установление зависимости непрерывной выходной переменной от входных переменных;

б) эта группировка объектов (наблюдений, событий) на основе данных, описывающих свойства объектов;

в) выявление закономерностей между связанными событиями;

г) это установление зависимости дискретной выходной переменной от входных переменных.

1.14. Установите соответствие между типами баз данных и их характеристиками:

1. Традиционная база данных	a. объем информации от петабайт до эксабайт
	b. централизованный способ хранения
	c. данные полуструктурированы и неструктурированы
2. Big Data	d. вертикальная модель хранения и обработки данных

	e. слабая взаимосвязь данных

1.15. Установите соответствие между объемами рынка Big Data и его сегментами:

1. Оборудование	a. 22%
2. Программное обеспечение	b. 38%
3. Сервисные услуги	c. 40%

1.16. Установите соответствие между наиболее распространенными подходами обработки данных (ПО) и их характеристиками:

1. NoSQL	a. используется для реализации поисковых и контекстных механизмов высоконагруженных сайтов – Facebook, eBay, Amazon и др. Отличительной особенностью является то, что система защищена от выхода из строя любого из узлов кластера, так как каждый блок имеет, как минимум, одну копию данных на другом узле.
2. Hadoop	b. Включает в себя ряд подходов, направленных на реализацию базы данных, имеющих отличия от моделей, используемых в традиционных, реляционных СУБД. Их удобно использовать при постоянно меняющейся структуре данных. Например, для сбора и хранения информации в социальных сетях.

1.18. Распределите нижеприведенные технологии по степени востребованности при использовании Big Data (наиболее востребованная – 1 место, наименее – 5).

1. Im-memory платформы (SAP HANA, Oracle Exadata и др.)	
2. Log-file аналитические программы (Splunk, InTrust Dell и др.)	
3. Columnar платформы (1010data, Calpont, HP Vertica и др.)	
4. NoSQL платформы	
5. Hadoop/MapReduce	

1.19. Установите соответствие между направлениями проектов по аналитике Big Data и проектами, в которых они используются:

1. Рекомендательные системы	a. компании во всех видах индустрии всегда занимались мониторингом и определением эффективности таких кампаний, но только возможности больших данных позволили получить анализ высокогранулированных потоков кликов на сайтах и звонков на телефонах с учетом местоположения клиентов.
2. Моделирование рисков	b. он-лайн магазины и Web – порталы используют проекты по аналитике Big Data, чтобы сравнивать пользовательские запросы и покупки, выделить поведенческие профили и характеристики и рекомендовать клиентам подходящие товары и услуги.
3. Анализ маркетинговых кампаний	c. финансовые компании, банки и некоторые другие используют проекты по аналитике Big Data и аналитическую песочницу, позволяющие анализировать большие объемы данных о транзакциях, определить риск и опасность финансовых активов, подготовить сценарии “что если”, основанные на имитационном моделировании поведения рынков, а также оценить

	риск от потенциальных клиентов

1.20. Ассоциация — …
а) это установление зависимости непрерывной выходной переменной от входных переменных;
б) эта группировка объектов (наблюдений, событий) на основе данных, описывающих свойства объектов;
в) выявление закономерностей между связанными событиями;
г) это установление зависимости дискретной выходной переменной от входных переменных.
1.21. Машинное обучение — …
а) специализированное программное решение (или набор решений), которое включает в себя все инструменты для извлечения закономерностей из сырых данных;
б) это группировка объектов (наблюдений, событий) на основе данных, описывающих свойства объектов;
в) набор данных, каждая запись которого представляет собой учебный пример, содержащий заданное входное влияние, и соответствующий ему правильный выходной результат;
г) подразделение искусственного интеллекта, изучающее методы построения алгоритмов, способных обучаться на данных.
1.22. Обучающая выборка — …
а) это группировка объектов (наблюдений, событий) на основе данных, описывающих свойства объектов;
б) набор данных, каждая запись которого представляет собой учебный пример, содержащий заданное входное влияние, и соответствующий ему правильный выходной результат;
в) набор данных, каждая запись которого представляет собой учебный пример, содержащий заданный входной влияние, что и обеспечивает ему правильный выходной результат;
г) выявление в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.
1.23. Установите соответствие между признаками Big Data и их характеристиками:

1. Volume	a. возможность одновременной обработки структурированной и неструктурированной разноформатной информации;
2. Velocity	b. накопленная база данных представляет собой большой объем информации, который трудоемко обрабатывать и хранить традиционными способами, для них требуются новый подход и усовершенствованные инструменты;
3. Variety	c. данный признак указывает как на увеличивающуюся скорость накопления данных (90% информации было собрано за последние 2 года), так и на скорость обработки данных, в последнее время стали более востребованы технологии обработки данных в реальном времени.

1.24. Эксперт это …
а) специалист в области анализа и моделирование;
б) специалист в предметной области;
в) человек, решать определенные задачи;
г) человек, который имеет опыт в программировании.
1.25. Виды лингвистической неопределенности:
а) неточность измерений значений определенной величины, выполняемых физическими приборами;
б) неопределенность значений слов (многозначность, размытость, непонятность, нечеткость); неоднозначность смысла фраз (синтаксическая и семантическая);
в) случайность (или наличие во внешней среде нескольких возможностей, каждая из которых случайным образом может стать действительностью); неопределенность значений слов (многозначность, размытость, неясность, нечеткость);
г) неоднозначность смысла фраз (синтаксическая и семантическая).
1.26. Установите соответствие между сферами применения Big Data и их процентным распределением:

1. Клиентский сервис	a. 7%

2. Операционная эффективность	b. 40%
3. Риск-менеджмент	c. 53%

1.27. Установите соответствие между драйверами и ограничителями рынка Big Data:

1. Драйверы	a. развитие облачной инфраструктуры;
	b. вопросы безопасности;
	c. персонал для внедрения проектов;
2. Ограничители	d. бюджет;
	e. изменения в законах о конфиденциальности данных;
	f. интеграция с существующими системами.

1.28. Распределите нижеприведенные страны по объему внедрения Big Data (страна с наибольшим объемом информации – 1 место, с наименьшим – 6):

1. Китай	
2. Бразилия	
3. Германия	
4. Индия	
5. Япония	
6. Германия	

1.29. Укажите фактор, способствовавший появлению тренда больших данных

а) маркетинговые кампании крупных корпораций;

б) снижение издержек на хранение данных;

в) появление новых технологий обработки потоковых данных;

г) выпуск баз данных с обработкой данных в памяти.

1.30. Отметьте значимые события, повлиявшие на формирование тренда больших данных:

а) разработка Hadoop;

б) изобретение принципа MapReduce;

в) разработка языка Pyhton;

г) победа Deepblue в матче с Г.Каспаровым.

1.31. Выберите верный ответ:

а) большие данные – это обработка или хранение более 1 Тб информации;

б) проблема больших данных – это такая проблема, когда при существующих технологиях хранения и обработки сущностная обработка данных затруднена или невозможна;

в) большие данные – это огромная PR-акция крупных вендоров и не более того;

г) большие данные – это явление, когда цифровые данные наиболее полно представляют изучаемый объект.

1.32. Выберите неверный ответ:

а) большие данные – это данные объёма свыше 1 Тб;

б) проблема больших данных – это проблема, когда при существующих технологиях хранения и обработки сущностная обработка данных затруднена или невозможна;

в) большие данные – это тренд в области ИТ, подогреваемый маркетинговыми кампаниями крупных вендоров;

г) большие данные как правило не структурированы.

1.33. Отметьте те из вариантов, в которых данные структурированы:

а) данные о продажах компании, представленные в виде помесячных отчётов в формате MS Word;

б) таблица с ежедневными показаниями температуры помещения за год в файле формата csv;

в) текст педагогической поэмы А.С. Макаренко, представленный в формате PDF;

г) библиотека фильмов, представленных в формате mpeg4 на одном жестком диске.

1.34. Перечислите четыре основных характеристики Big Data:

а) Virtualization, Volume, Variability, Vehicle;

б) Variety, Velocity, Volume, Value;

в) Verification, Volume, Velocity, Visualization;

г) Video, Value, Variety, Volume.

1.35. Выберите неверное высказывание:

а) большие объёмы данных приводят к слабой их структуризации, поэтому появляется такое разнообразие данных;

б) увеличившаяся производительность телекоммуникационных каналов привела к росту объёмов передаваемой информации;

в) удешевление систем хранения на единицу информации привело к росту рынка больших данных;
г) большое разнообразие источников данных
1.36. Отметьте неверное понимание Variety в контексте характеристик Big Data:
а) высокая скорость генерирования данных;
б) разные типы данных в колонках таблиц реляционных СУБД;
в) разнообразие отраслей, являющихся источниками данных;
г) разнообразие типов данных, включающих в себя структурированные, полуструктурированные и неструктурированные.
1.37. Принцип MapReduce состоит в том, чтобы:
а) производить вычисления на узлах, где информация изначально была сохранена;
б) использовать вычислительные мощности систем хранения;
в) использовать функциональное программирование для решения задач массивно-параллельной обработки.
1.38. Выберите одно неверное высказывание про MapReduce:
а) интерфейс для массово-параллельной обработки данных, где вычисления производятся на узлах, где информация изначально была сохранена;
б) MapReduce – это две операции: распределения и сборки данных;
в) MapReduce был придуман разработчиками Hadoop;
г) MapReduce был анонсирован разработчиками Google.
1.39. Какие из следующих технологий СУБД не используют принцип MapReduce:
а) Hadoop;
б) Cassandra;
в) HDInsight;
г) Redis.
1.40. Какие СУБД полностью полагаются на оперативную память при хранении информации:
а) Oracle Exalytics;
б) SAP HANA;
в) BigTable;
г) HBase.
1.41. В чём преимущество колоночно-ориентированных СУБД?
а) они позволяют выполнять более сложные SQL-запросы по сравнению с реляционными СУБД;
б) они позволяют динамически дополнять содержание записей новыми полями;
в) они имеют более гибкие возможности аналитики;
г) они позволяют эффективно делать межколоночные сравнения.
1.42. Расставьте последовательность этапов проекта аналитики в соответствии с CRISP-DM:
а) понимание бизнеса (Business understanding);
б) понимание данных (Data Understanding);
в) подготовка данных (Data Preparation);
г) моделирование (Modeling);
д) оценка (Evaluation);
е) внедрение (Deployment).
1.43. На каком из этапов процесса CRISP-DM происходит проверка гипотез?
а) понимание бизнеса (Business understanding);
б) понимание данных (Data Understanding);
в) моделирование (Modeling);
г) оценка (Evaluation).
1.44. Вы являетесь владельцем и аналитиком в компании из 10 человек, в которой требуется проанализировать продажи за 1 год (1 млн. продаж). Какие из этапов CRISP-DM можно опустить:
а) понимание бизнеса (Business understanding);
б) подготовка данных (Data Preparation);
в) моделирование (Modeling);
г) оценка (Evaluation).
1.45. Пример благоразумного использования Hadoop:
а) анализ 10 Гб данных;
б) ежедневное сохранение данных температуры, поступающих со всех городов России (по одному показанию на город, всего городов 1100 шт.);
в) посекундное сохранение данных температуры, поступающих со всех городов России (по одному

показанию на город, всего городов 1100 шт.);
г) построение графика пульса пациента в реальном времени.
1.46. Начиная с каких размеров данных обоснованно применение кластера Hadoop для хранения данных?
а) 100Гб;
б) 1Тб;
в) 100Тб;
г) 1Пб.
1.47. Hadoop – это:
а) набор утилит, и программный каркас для выполнения распределённых программ, работающих на кластерах;
б) распределённая СУБД, позволяющая обрабатывать большие данные;
в) язык выполнения заданий в парадигме MapReduce;
г) распределённая файловая система, предназначенная для хранения файлов большого объёма.

Модуль 2 «Методы и модели анализа больших данных»

2.1. К описательным моделям относятся следующие модели данных:
а) модели классификации и последовательностей;
б) регрессионные, кластеризации, исключений, итоговые и ассоциации;
в) классификации, кластеризации, исключений, итоговые и ассоциации;
г) модели классификации, последовательностей и исключений.
2.2. Модели классификации описывают …
а) правила или набор правил, в соответствии с которыми можно отнести описание любого нового объекта к одному из классов;
б) функции, которые позволяют прогнозировать изменения непрерывных числовых параметров;
в) функциональные зависимости между зависимыми и независимыми показателями и переменными в понятной человеку форме;
г) группы, на которые можно разделить объекты, данные о которых подвергаются анализу.
2.3. Модели исключений описывают …
а) исключительные ситуации в записях, которые резко отличаются по произвольному признаку от основного множества записей;
б) ограничения на данные анализируемого массива;
в) закономерности между связанными событиями;
г) группы, на которые можно разделить объекты, данные о которых подвергаются анализу.
2.4. Итоговые модели обнаружат …
а) исключительные ситуации в записях, которые резко отличаются по произвольному признаку от основного множества записей;
б) ограничения на данные анализируемого массива;
в) закономерности между связанными событиями;
г) группы, на которые можно разделить объекты, данные о которых подвергаются анализу.
2.5. Установите соответствие между алгоритмами объединения двух кластеров и их характеристиками:

1. Метод дальнего соседа	a. Степень близости оценивается как средняя величина степеней близости между объектами кластеров.
2. Метод средней связи	b. Расстояние между любым кластером S и новым кластером, который получился в результате объединения кластеров P и Q, определяется как расстояние от центра кластера S до середины отрезка, соединяющего центры кластеров P и Q.
3. Метод медианной связи	c. Степень близости оценивается по степени близости между наиболее отдаленными объектами кластеров

2.6. Задача регрессии сводится к …
а) нахождению частных зависимостей между объектами или событиями;
б) определения класса объекта по его характеристикам;
в) определение по известным характеристикам объекта значение некоторого его параметра;
г) поиска независимых групп и их характеристик во всем множестве анализируемых данных.

2.7. Модели последовательностей описывают …
а) правила или набор правил, в соответствии с которыми можно отнести описание любого нового объекта к одному из классов;
б) функции, которые позволяют прогнозировать изменения непрерывных числовых параметров;
в) функциональные зависимости между зависимыми и независимыми показателями и переменными в понятной человеку форме;
г) группы, на которые можно разделить объекты, данные о которых подвергаются анализу.
2.8. Модели ассоциации проявляют …
а) исключительные ситуации в записях, которые резко отличаются по произвольному признаку от основной множества записей;
б) ограничения на данные анализируемого массива;
в) закономерности между связанными событиями;
г) группы, на которые можно разделить объекты, данные о которых подвергаются анализу.
2.9. Регрессионные модели описывают …
а) правила или набор правил в соответствии с которыми можно отнести описание любого нового объекта к одному из классов;
б) функции, которые позволяют прогнозировать изменения непрерывных числовых параметров;
в) функциональные зависимости между зависимыми и независимыми показателями и переменными в понятной человеку форме;
г) группы, на которые можно разделить объекты, данные о которых подвергаются анализа.
2.10. Задача кластеризации заключается в …
а) нахождении частых зависимостей между объектами или событиями;
б) определении класса объекта по его характеристикам;
в) определении по известным характеристикам объекта значение некоторого его параметра;
г) поиске независимых групп и их характеристик во всем множестве анализируемых данных.
2.11. Ошибка обучения — …
а) это ошибка, допущенная моделью на учебном множестве;
б) это ошибка, полученная на тестовых примерах, то есть, что вычисляется по тем же формулам, но для тестового множества;
в) имена, типы, метки и назначения полей исходной выборки данных;
г) набор данных, каждая запись которого представляет собой учебный пример, содержащий заданное входное влияние, и соответствующий ему правильный выходной результат.
2.12. Ошибка обобщения — …
а) это ошибка, допущенная моделью на учебном множестве;
б) это ошибка, полученная на тестовых примерах, то есть, вычисляется по тем же формулам, но для тестового множества;
в) имена, типы, метки и назначения полей исходной выборки данных;
г) набор данных, каждая запись которого представляет собой учебный пример, содержащий заданное входное влияние, и соответствующий ему правильный выходной результат.
2.13. Аналитик это …
а) специалист в области анализа и моделирования;
б) специалист в предметной области;
в) человек, решающий определенные задачи;
г) человек, который имеет опыт в программировании.
2.14. Задача классификации сводится к …
а) нахождению частых зависимостей между объектами или событиями;
б) определению класса объекта по его характеристикам;
в) определению по известным характеристикам объекта значение некоторого его параметра;
г) поиску независимых групп и их характеристик в всем множестве анализируемых данных.
2.15. Установите соответствие между направлениями проектов по аналитике Big Data и проектами, в которых они используются:

1. Анализ “чувств” (sentiment analysis)	а) использование техник больших данных для объединения поведения клиентов, исторических данных и транзакций, чтобы детектировать активности, подозрительные на хищения. Компании кредитных карт используют технологии

	больших данных для определения по потоку транзакций подозрительных на операции с похищенной карты;
2. Детектирование хищений	b) благодаря новым видам хранилищ данных для эффективного хранения больших графов и Big Data из структур связей между людьми извлекается информация о степени влияния отдельных персон на другие. Это помогает выделить “наиболее важных” клиентов, которыми часто являются вовсе не те, кто покупает больше всего продуктов, а те, чье мнение влияет на большинство других клиентов компании;
3. Анализ социальных графов	c) использован расширенная аналитика текста на основе Big Data на основе анализа неструктурированных текстов в социальных сетях и СМИ включая посты в Twitter и Facebook, чтобы определить какие чувства испытывает пользователь к выбранной компании, бренду или продукту. Анализ может фокусироваться как на статистически средние типы людей, так и индивидуальных пользователей.

2.16. «Песочница» в аналитическом процессе:
а) для чего аналитику необходима «песочница»?
б) для высокопроизводительной аналитики за счёт использования оперативной памяти и inDB операций;
в) для хранения всех полученных от заказчика данных;
г) для построения отчётов о результатах анализа;
д) для снижения затрат, связанных с репликацией данных.
2.18. Какие из следующих средств разумно использовать для анализа данных, представленных единственным csv-файлом размера более 100Гб:
а) Hadoop;
б) Data Warehouse;
в) «Песочница»;
г) Python.
2.19. Выберите верное утверждение:
а) Data Warehouse создаются для проверки гипотез при анализе больших данных;
б) «Песочница» используется для снижения нагрузки на основной Data Warehouse;
в) каждый Data Warehouse должен содержать «песочницу»;
г) «Песочница» необходима для любого процесса аналитики.

Критерии оценивания:
Максимальный балл – 20.

Число вопросов - 20. Ответ на каждый вопрос оценивается максимум в 1 балл.

Критерии оценивания 1 вопроса:

0,84-1,0 балла выставляется студенту, если изложенный материал фактически верен, продемонстрированы глубокие исчерпывающие знания в объеме пройденной программы в соответствии с поставленными программой курса целями и задачами обучения, изложение материала при ответе грамотное и логически стройное;

0,67-0,83 балла выставляется студенту, если продемонстрированы твердые и достаточно полные знания в объеме пройденной программы дисциплины в соответствии с целями обучения; материал изложен достаточно полно с отдельными логическими и стилистическими погрешностями;

0,5-0,66 балла выставляется студенту, если продемонстрированы твердые знания в объеме пройденного курса в соответствие с целями обучения, ответ содержит отдельные ошибки, уверенно исправленные после дополнительных вопросов;

0-0,49 балла выставляется студенту, если ответ не связан с вопросом, допущены грубые

ошибки в ответе, продемонстрированы непонимание сущности излагаемого вопроса, неуверенность и неточность ответов на дополнительные и наводящие вопросы.

Вопросы для собеседования

Модуль 1 «Сбор, хранение и анализ больших данных»

1. Что такое Big Data? Понятие, сущность и ключевые признаки больших данных.
2. Роль и место больших данных в решении аналитических и исследовательских задач профессиональной деятельности.
3. Методы и средства анализа Big Data.
4. Используемые программы для анализа Big Data.
5. Технологии хранения больших данных.
6. Определение источника больших данных.
7. Исследование источника данных.
8. Что такое хранилище данных?
9. Процесс анализа больших данных.
10. Технологии анализа больших данных.
11. Научные проблемы в области больших данных.
12. Технологии и инструменты больших данных.
13. Apache Hadoop. Storm – система потоковой обработки.
14. Язык программирования R.
15. Аналитика больших данных как корпоративный проект.
16. Сущность и принцип работы аналитической платформы Deductor Academic.
17. Основные функции и инструменты аналитической платформы Deductor Academic для целей анализа и исследования социально-экономических процессов и явлений в деятель-ности предприятий.
18. Моделирование социально-экономических процессов и явлений в деятельности пред-приятий с помощью платформы Deductor Academic.
19. Инструментарий прикладного компьютерного анализа и моделирования в Deductor Academic.

Модуль 2 «Методы и модели анализа больших данных»

20. Прогнозирование и предвидение в социально-политических и медиа процессах.
21. Методы прогнозирования, использующие большие данные.
22. Что такое OLAP-система?
23. Чем OLAP-системы отличаются от Big Data?
24. Техники больших данных.
25. Консолидация данных.
26. Визуализация.
27. Классификация.
28. Кластеризация.
29. Регрессионный анализ.
30. Анализ ассотиативных правил.
31. Нейронные сети. Экспоненциальное сглаживание Холта с корректировкой тренда в моделях временных рядов.
32. Интеллектуальный анализ данных (Data Mining).
33. Задачи и методы интеллектуального анализа больших данных.
34. Инструменты Data Mining.

Общее число вопросов на собеседовании 3. Каждый вопрос оценивается отдельно, максимально в 5 балла.

Критерии оценивания отдельного вопроса:

- 2,5-5,4 балла. Ответ на вопрос верный; продемонстрировано наличие глубоких исчерпывающих / твердых и достаточно полных знаний, грамотное и логически стройное изложение материала при ответе, возможны отдельные погрешности и ошибки, уверенно исправленные и после дополнительных вопросов; продемонстрировано наличие глубоких исчерпывающих / твердых и достаточно полных знаний, грамотное и логически стройное изложение материала при ответе.
- 0-2,4 балла. Ответ на вопрос лишь частично верен, продемонстрирована неточность и неуверенность ответов на дополнительные и наводящие вопросы, либо ответ на вопрос не верен, продемонстрирована неуверенность и неточность ответов на дополнительные и наводящие вопросы.

Комплект разноуровневых задач (заданий)

1 Задачи репродуктивного уровня

Задача 1. Исследователь анализирует зависимость потребления (c) от располагаемого дохода (y) на основе простой эмпирической модели: $c_i=\beta y_i+\varepsilon_i$, ε_i - независимые нормально распределенные случайные величины с нулевым математическим ожиданием и дисперсией $V(\varepsilon_i) = a^2 \cdot y_i^2$.

Исследователь собрал данные о двух тысячах домашних хозяйств и осуществил следующие предварительные расчёты:

$\sum_{i=1}^{2000} y_i = 2000;\ \sum_{i=1}^{2000} c_i = 1000;\ \sum_{i=1}^{2000} y_i^2 = 1450;\ \sum_{i=1}^{2000} y_i c_i = 950;\ \sum_{i=1}^{2000} \frac{y_i}{c_i} = 1050;\ \sum_{i=1}^{2000} \frac{c_i}{y_i} = 1550.$

Используя те из доступных данных, которые вам необходимы, вычислите эффективную оценку предельной склонности к потреблению.

Задача 2. Исследуется зависимость среднедушевого потребления алкоголя по странам мира от различных факторов.

Модель 1:

$$ALCO_i=\beta_1+\beta_2 GDP_i+\beta_3 MUSL_i+\beta_4 BUDD_i+\beta_5 HINDU_i+\varepsilon_i,$$

где $ALCO_i$ — среднедушевое потребление чистого спирта на человека (л), GDP_i — ВВП на душу населения (долларов США), $MUSL_i$, $BUDD_i$, $HINDU_i$ — доли населения исповедующего, соответственно, мусульманство, буддизм и индуизм (в % от общей численности населения). В ходе МНК-оценивания модели на основе данных о 50 странах получены следующие результаты: сумма квадратов остатков $ESS=200$, объясненная сумма квадратов $RSS=300$.

Также для проверки гипотезы о том, что религия не оказывает существенного влияния на потребление алкоголя, были оценены параметры второй модели:

Модель №2:

$$ALCO_i=\beta_1+\beta_2 GDP_i+\varepsilon_i.$$

Во второй модели, по сравнению с первой, значение *RSS* изменилось на100. Сколько составит корректированный R^2 во второй модели?

Задача 3. В переменной df сохранен subset данных mtcars только с переменными "wt", "mpg", "disp", "drat", "hp". Воспользуйтесь множественным регрессионным анализом, чтобы предсказать вес машины (переменная "wt"). Выберите такую комбинацию независимых переменных (из "mpg", "disp", "drat", "hp"), чтобы значение R^2 adjusted было наибольшим. Взаимодействия факторов учитывать не надо.

2 Задачи реконструктивного уровня

Задача 1. Исходный файл с данными: EARNINGS.xls

В вашем распоряжении имеются следующие данные о 540 работниках (270 мужчин и 270 женщин): *EARNINGS* — текущий часовой заработок в долларах США, *S* — продолжительность обучения(число полных лет обучения), *EXP* — общий стаж работы после окончания учебы, *FEMALE* — пол респондента(0 — для мужчин, 1 — для женщин).

Импортируйте данные в R.

Постройте модель вида: $EARNINGS_i=\beta_1+\beta_2+\beta_3 EXP_i+\beta_4 FEMALE_i+\varepsilon_i$

Проведя тест Бреуша-Пагана, скажите, присутствует ли в модели гетероскедастичность, и если да, то с какой переменной она, скорее всего, связана (при ответе ориентируйтесь на значимость коэффициентов в соответствующем уравнении).

Задача 2. Проанализируйте факторы, которые влияют на выручку компании Freeny. Для этого подгрузите встроенный в R массив данных *freeny*. Оцените зависимость выручки $lag.quarterly.revenue_i$ от индекса цен $price.index_i$ и дохода $income.level_i$

$$lag.quarterly.revenue_i = \beta 1 + \beta 2 price.index_i + \beta 3 income.level_i + \varepsilon_i$$

К какому выводу Вы придете, проверяя гипотезу о значимости индекса цен на уровне значимости 5%.

Задача 3. Исследователь анализирует влияние вступления в брак на уровень доходов. Он собрал данные за четыре года о тысяче работников обоих полов, часть из которых в течение рассматриваемого периода вступили в брак. В качестве объясняющей исследователь использовал фиктивную переменную *Одинокий*, которая равна единице для тех работников, которые в данном году не женаты (не замужем). В качестве контрольных переменных он использовал переменные *Возраст* (возраст в годах) и *Образование* (число лет обучения в годах). Исследователь оценил четыре уравнения (см. таблицу): первые два оценены обычным МНК, третье и четвертое оценены при помощи модели с фиксированными эффектами (внутригрупповое преобразование).

Таблица 1. Результаты оценки моделей. Зависимая переменная — логарифм заработной платы работника.

Модель	**Модель 1**	**Модель 2**	**Модель 3**	**Модель 4**
Метод оценивания	МНК	МНК	FE	FE
Возраст	0,80 (0,10)	0,78 (0,09)	0,67 (0,11)	0,68 (0,11)
Число лет обучения	1,20 (0,24)	1,41 (0,23)	0,91 (0,50)	0,90 (0,52)
Одинокий	–0,05 (0,02)	–0,04 (0,02)	–0,02 (0,01)	–0,03 (0,01)
Индивидуальные эффекты	Нет	Нет	Да	Да
Фиктивные переменные времени	Нет	Да	Нет	Да
Число наблюдений	4000	4000	4000	4000
R-квадрат	0,612	0,723	0,520	0,521
P-значение теста на отсутствие индивидуальных эффектов	—	—	0,001	0,002
P-значение теста на равенство нулю коэффициентов при фиктивных переменных времени	—	0,090	—	0,170

В скобках под оценками коэффициентов указаны робастные стандартные ошибки. В случае МНК представлен скорректированный R-квадрат, в случае модели с фиксированными эффектами within-R-квадрат.

Какую из четырех моделей следует выбрать в соответствии с доступной информацией?

3 Задачи творческого уровня

Задача 1. Используя открытые статистические данные, постройте регрессионную модель на больших данных. Приведите теоретическое обоснование вида и структуры модели. Оцените ее качество. Сделайте выводы.

Задача 2. Проведите исследование, чтобы выявить факторы, влияющие на доход людей, на реальных данных Российского мониторинга экономического положения и здоровья населения (РМЭЗ, он же RLMS). Воспользуйтесь для этого репрезентативной выборкой по индивидам за 2014 год - волна 23 (нужный файл называется r23i_os26a.sav). Для загрузки данных в R воспользуйтесь пакетом rlms.

Набор включает по одной задаче каждого уровня (суммарно максимально 10 баллов).
Задача репродуктивного уровня оценивается максимально в 2 балла:
Критерии оценивания задачи:

- 1-2 балла. Задача решена в полном объеме, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов; либо задача решена в полном объеме с небольшими погрешностями, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов, в расчетах и выводах содержатся незначительные ошибки.
- 0-0,9 балла. Задача решена частично, частично выбраны верные инструментальные методы и приемы решения, проведены частичные расчеты, сделан вывод по результатам проведенных расчетов с погрешностями либо задача не решена или решена частично, частично выбраны необходимые инструментальные методы и приемы решения, расчеты не проведены или проведены частично, вывод по результатам проведенных расчетов не сделан или ошибочен.

Задача реконструктивного уровня оценивается максимально в 3 балла:
Критерии оценивания задачи:

- 1,5-3 балла. Задача решена в полном объеме, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов; либо задача решена в полном объеме с небольшими погрешностями, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов, в расчетах и выводах содержатся незначительные ошибки.
- 0-1,4 балла. Задача решена частично, частично выбраны верные инструментальные методы и приемы решения, проведены частичные расчеты, сделан вывод по результатам проведенных расчетов с погрешностями либо задача не решена или решена частично, частично выбраны необходимые инструментальные методы и приемы решения, расчеты не проведены или проведены частично, вывод по результатам проведенных расчетов не сделан или ошибочен.

Задача творческого уровня оценивается максимально в 5 баллов:
Критерии оценивания задачи:

- 2,5-5 баллов. Задача решена в полном объеме, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов; либо задача решена в полном объеме с небольшими погрешностями, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов, в расчетах и выводах содержатся незначительные ошибки.
- 0-2,4 баллов. Задача решена частично, частично выбраны верные инструментальные методы и приемы решения, проведены частичные расчеты, сделан вывод по результатам проведенных расчетов с погрешностями либо задача не решена или решена частично, частично выбраны необходимые инструментальные методы и приемы решения, расчеты не проведены или проведены частично, вывод по результатам проведенных расчетов не сделан или ошибочен.

Лабораторные работы

1.Тематика лабораторных работ по разделам и темам

Модуль 1 «Сбор, хранение и анализ больших данных»

1. Введение в R. Работа с электронными таблицами больших данных в Calc.
2. R: визуализация Big Data, фиктивные переменные, прогнозы, проверка гипотез и ловушка дамми-переменных.
3. Большие данные и скрипты. Проблемы регрессионного анализа больших данных (мультиколлинеарность и гетероскедастичность) в R.

Модуль 2 «Методы и модели анализа больших данных»

4. Прогнозное моделирование: работа с регрессионными моделями больших данных в Calc и R.
5. R: даты и временные ряды, загрузка больших данных и тесты на автокорреляцию, качественные переменные, предельные эффекты и ROC кривая. Экспоненциальное сглаживание Холта с корректировкой тренда в моделях временных рядов больших данных в Calc и R.
6. Эконометрическое исследование на больших данных Российского мониторинга экономического положения и здоровья населения (РМЭЗ, он же RLMS).
7. Кластерный анализ на больших данных. Анализ потребительской корзины. Использование метода k-средних для сегментирования клиентской базы. Сетевые графы и определение сообществ.
8. Примеры анализа временных рядов больших данных в R. Определение выбросов.

Критерии оценивания:
Максимальная сумма баллов за все лабораторные работы = 55 баллов.
Каждая лабораторная работа №1-3 оценивается максимально в 5 баллов.
Критерии оценки каждой работы:

- 2,5-5 баллов. Задание выполнено в полном объеме, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов; либо задание решено в полном объеме с небольшими погрешностями, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов, в расчетах и выводах содержатся незначительные ошибки.
- 0-2,4 баллов. Задание выполнено частично, частично выбраны верные инструментальные методы и приемы решения, проведены частичные расчеты, сделан вывод по результатам проведенных расчетов с погрешностями либо задание не выполнено или выполнено частично, частично выбраны необходимые инструментальные методы и приемы решения, расчеты не проведены или проведены частично, вывод по результатам проведенных расчетов не сделан или ошибочен.

Каждая лабораторная работа №4-8 оценивается максимально в 8 баллов.
Критерии оценки каждой работы:

- 4,1-8 баллов. Задание выполнено в полном объеме, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов; либо задание решено в полном объеме с небольшими погрешностями, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов, в расчетах и выводах содержатся незначительные ошибки.
- 0-4 балла. Задание выполнено частично, частично выбраны верные инструментальные методы и приемы решения, проведены частичные расчеты, сделан вывод по результатам проведенных расчетов с погрешностями либо задание не выполнено или выполнено частично, частично выбраны необходимые инструментальные методы и приемы решения, расчеты не проведены или проведены частично, вывод по результатам проведенных расчетов не сделан или ошибочен.

3. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций

Процедуры оценивания включают в себя текущий контроль и промежуточную аттестацию.

Текущий контроль успеваемости проводится с использованием оценочных средств, представленных в п. 2 данного приложения. Результаты текущего контроля доводятся до сведения студентов до промежуточной аттестации.

Промежуточная аттестация проводится в форме зачета.

Зачет проводится по расписанию промежуточной аттестации в письменном виде. В задании к зачету – 2 теоретических вопроса и 1 задача. Проверка ответов и объявление результатов производится в день зачета. Результаты аттестации заносятся в зачетную ведомость и зачетную книжку студента. Студенты, не прошедшие промежуточную аттестацию по графику сессии, должны ликвидировать задолженность в установленном порядке.

Приложение 2

МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ

Методические указания адресованы студентам очной формы обучения.

Учебным планом предусмотрены следующие виды занятий:

- практические занятия;
- лабораторные работы.

В ходе практических занятий рассматриваются методы анализа и синтеза в предметной области; современные методы анализа данных; возможные ограничения применения статистических и эконометрических методов; методики совершенствования знаний в области анализа Big Data, даются рекомендации по самостоятельной работе и подготовке к лабораторным работам.

В ходе лабораторных работ углубляются и закрепляются знания студентов по ряду рассмотренных на практических занятиях вопросов, формируются и развиваются навыки использовать современное программное обеспечение (RStudio) для решения экономико-статистических и эконометрических задач обработки данных: построение таблиц, визуализация, проверка гипотез, корреляционно-регрессионный анализ, анализ временных рядов и панельных данных, анализ текстовой и графической информации.

При подготовке к лабораторным работам каждый студент должен:

– изучить рекомендованную учебную литературу;

– подготовить ответы на все вопросы по изучаемой теме.

В процессе подготовки к лабораторным работам студенты могут воспользоваться консультациями преподавателя.

Вопросы, не рассмотренные на практических занятиях и лабораторных работах, должны быть изучены студентами в ходе самостоятельной работы. Контроль самостоятельной работы студентов над учебной программой курса осуществляется в ходе занятий методом устного опроса или посредством тестирования. В ходе самостоятельной работы каждый студент обязан прочитать основную и по возможности дополнительную литературу по изучаемой теме. Выделить непонятные термины, найти их значение в энциклопедических словарях.

Для подготовки к занятиям, текущему контролю и промежуточной аттестации студенты могут воспользоваться электронно-библиотечными системами. Также обучающиеся могут взять на дом необходимую литературу на абонементе университетской библиотеки или воспользоваться читальными залами.