

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Новосибирский государственный технический университет»

Кафедра теоретической и прикладной информатики
Кафедра прикладной математики

“УТВЕРЖДАЮ”
ДЕКАН ФПИИ
д.т.н. Тимофеев В. С.
“ ____ ” _____ г.

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ
Статистические методы анализа данных

Образовательная программа: 01.03.02 Прикладная математика и информатика, профиль:
Компьютерное моделирование и информационные технологии

Курс: 4, семестр : 7

Факультет прикладной математики и информатики,

		Семестр
№	Вид деятельности	7
1	Всего зачетных единиц (кредитов)	3
2	Всего часов	108
3	Всего занятий в контактной форме, час.	46
4	Лекции, час.	18
5	Практические занятия, час.	0
6	Лабораторные занятия, час.	18
7	из них в активной и интерактивной форме, час.	5
8	Аттестация, час.	2
9	Консультации, час.	8
10	Самостоятельная работа, час.	62
11	Виды самостоятельной работы (курсовой проект, курсовая работа, РГЗ, подготовка к контрольной работе)	РГЗ
12	Вид аттестации	ДЗ

Рабочая программа составлена на основании федерального государственного образовательного стандарта высшего образования по направлению (специальности): 01.03.02 Прикладная математика и информатика

ФГОС введен в действие приказом №228 от 12.03.2015 г. , дата утверждения: 14.04.2015 г.

Место дисциплины в структуре учебного плана: Блок 1, вариативная, по выбору студента

Рабочая программа разработана на основе компетентностной модели выпускника по направлению (специальности): 01.03.02 Прикладная математика и информатика

Рабочая программа обсуждена на заседании кафедры

ТПИ, протокол заседания кафедры №4 от 20.06.2017

ПМт, протокол заседания кафедры №6 от 20.06.2017

Утверждена на совете факультета прикладной математики и информатики, протокол № 6 от 21.06.2017

Программу разработал:

профессор, д.т.н. Попов А. А.

Заведующий кафедрой:

доцент, д.т.н. Чубич В. М.

профессор, д.т.н. Соловейчик Ю. Г.

Ответственный за образовательную программу:

заведующий кафедрой Соловейчик Ю. Г.

1. Внешние требования

Таблица 1.1

Компетенция ФГОС: ОПК.1 способность использовать базовые знания естественных наук, математики и информатики, основные факты, концепции, принципы теорий, связанных с прикладной математикой и информатикой; в части следующих результатов обучения:	
35. знать методы теории вероятностей и математической статистики	
Компетенция ФГОС: ПК.2 способность понимать, совершенствовать и применять современный математический аппарат; в части следующих результатов обучения:	
34. знать теоретические основы методов построения зависимостей по экспериментальным данным	
у2. уметь применять основные математические методы при построении моделей	

2. Требования НГТУ к результатам освоения дисциплины

Таблица 2.1

Результаты изучения дисциплины по уровням освоения (иметь представление, знать, уметь, владеть)	Формы организации занятий
<i>Статистические методы анализа данных</i>	
ОПК.1.35 знать методы теории вероятностей и математической статистики	
1.Знать базовые методы анализа статистических данных при построении зависимостей	Лекции; Самостоятельная работа
ПК.2.34 знать теоретические основы методов построения зависимостей по экспериментальным данным	
2.знать теоретические основы методов построения зависимостей по экспериментальным данным	Лекции; Самостоятельная работа
ПК.2.у2 уметь применять основные математические методы при построении моделей	
3.Владеть методами построения и анализа линейных моделей по экспериментальным данным	Лабораторные работы; Самостоятельная работа

3. Содержание и структура учебной дисциплины

Таблица 3.1

Темы лекций	Активные формы, час.	Часы	Ссылки на результаты обучения	Учебная деятельность
Семестр: 7				
Дидактическая единица: Классификация задач статистического анализа данных. Задачи статистического исследования зависимостей, задачи снижения размерности, задачи классификации, задачи кластеризации. Методы интеллектуального анализа данных (Data Mining)				
1. Задачи статистического анализа данных.	0	2	1, 2	Учебная деятельность состоит в изучении теоретического материала, подготовки и выполнения лабораторных работ, их защита
Дидактическая единица: Параметрические регрессионные схемы. Основные требования, предъявляемые к оценкам параметров. Наилучшие линейные оценки (НЛО). Вычисление НЛО, их свойства. Связь НЛО с оценками по методу наименьших квадратов (МНК). Оценки по методу максимального правдоподобия				
2. Многомерный регрессионный анализ. Оценивание параметров	0	4	1, 2	Учебная деятельность состоит в изучении теоретического материала, подготовки и выполнения лабораторных работ, их защита

Дидактическая единица: Интервальное оценивание и проверка гипотез. Оценивание неизвестной дисперсии ошибки наблюдения. Проверка адекватности регрессионной модели. Оценивание параметров в модели регрессии при наличии линейных ограничений. Проверка общих линейных гипотез. Доверительное оценивание.				
3. Многомерный регрессионный анализ. Интервальное оценивание и проверка гипотез.	0	4	1, 2	Учебная деятельность состоит в изучении теоретического материала, подготовки и выполнения лабораторных работ, их защита
Дидактическая единица: Оценивание параметров в условиях гетероскедастичности, автокорреляции ошибок наблюдения, в условиях мультиколлинеарности				
4. Расширения общей линейной модели	0	4	1, 2	Учебная деятельность состоит в изучении теоретического материала, подготовки и выполнения лабораторных работ, их защита
Дидактическая единица: Выбор общего вида функции регрессии (структуры модели). Выбор наилучшей модели. Критерии качества модели. Алгоритмы выбора наилучшей модели регрессии: алгоритмы включения и исключения, эффективный алгоритм построения всех регрессий, многорядный комбинаторный алгоритм.				
5. Выбор общего вида функции регрессии	0	4	2	Учебная деятельность состоит в изучении теоретического материала, подготовки и выполнения лабораторных работ, их защита

Таблица 3.2

Темы лабораторных работ	Активные формы, час.	Часы	Ссылки на результаты обучения	Учебная деятельность
Семестр: 7				
Дидактическая единица: Параметрические регрессионные схемы. Основные требования, предъявляемые к оценкам параметров. Наилучшие линейные оценки (НЛО). Вычисление НЛО, их свойства. Связь НЛО с оценками по методу наименьших квадратов (МНК). Оценивание по методу максимального правдоподобия				
1. Моделирование регрессионных процессов	0	2	3	В соответствии с вариантом задания осуществляет выбор имитационной модели объекта, диапазон изменения факторов, план эксперимента. Разрабатывает программу имитационного моделирования для генерации экспериментальных данных. Подготавливает полученные по программе данные для дальнейшей обработки. Осуществляет построение графиков зависимости отклика от входных факторов. Оформляет отчет

2. Оценивание неизвестных параметров линейных регрессионных моделей	1	4	3	Осуществляет проектирование и формирование программного модуля по вычислению МНК-оценок параметров для заданной параметрической модели объекта. Пользуясь экспериментальными данными, полученными в лабораторной работе № 1, производит оценку параметров модели объекта. Проводит проверку адекватности полученной модели. Оформляет отчет, включающий постановочную часть, текст программы, листинг вычислений, принятые решения по проверке адекватности модели
Дидактическая единица: Интервальное оценивание и проверка гипотез. Оценивание неизвестной дисперсии ошибки наблюдения. Проверка адекватности регрессионной модели. Оценивание параметров в модели регрессии при наличии линейных ограничений. Проверка общих линейных гипотез. Доверительное оценивание.				
3. Интервальное оценивание неизвестных параметров регрессионных, моделей проверка гипотез	1	4	3	По полученным в лабораторной работе №1 экспериментальным данным осуществляет построение доверительных интервалов для каждого параметра модели регрессии. Проводит проверку гипотезы о незначимости каждого параметра модели, гипотезы о незначимости самой регрессии. Осуществляет построение графика прогнозных значений и доверительной полосы для математического ожидания функции отклика и для самого отклика. Оформляет отчет, включающий постановочную часть, текст программы, листинг вычислений
Дидактическая единица: Оценивание параметров в условиях гетероскедастичности, автокорреляции ошибок наблюдения, в условиях мультиколлинеарности				

4. Оценивание неизвестных параметров регрессионных моделей в условиях гетероскедастичности возмущений	1	2	3	Осуществляет моделирование регрессионного процесса с гетероскедастичным возмущением. Полученные данные проверяет по тестам на наличие гетероскедастичности. Осуществляет оценивание параметров регрессионной модели по доступному обобщенному МНК и по обыкновенному МНК. Проводит сравнение эффективности оценок в этих двух случаях по квадрату их расстояния до известных истинных значений параметров. Оформляет отчет, включающий постановочную часть, текст программы, листинг вычислений
5. Оценивание неизвестных параметров регрессионных моделей в условиях мультиколлинеарности	1	2	3	Осуществляет генерацию экспериментальных данных, в которых присутствует эффект мультиколлинеарности. Рассчитывает ряд показателей, характеризующих эффект мультиколлинеарности. Определяет факторы, ответственные за возникновение эффекта мультиколлинеарности. Вычисляет ридж-оценки параметров при различных значениях параметра регуляризации. Выбирает оптимальное значение параметра регуляризации. Проводит оценивание модели регрессии по методу главных компонентов. Сравнивает решение с ридж-оцениванием по смещению оценок и точности предсказания отклика. Оформляет отчет, включающий постановочную часть, текст программы, листинг вычислений
Дидактическая единица: Выбор общего вида функции регрессии (структуры модели). Выбор наилучшей модели. Критерии качества модели. Алгоритмы выбора наилучшей модели регрессии: алгоритмы включения и исключения, эффективный алгоритм построения всех регрессий, многорядный комбинаторный алгоритм.				

6. Последовательные алгоритмы поиска модели объекта	1	2	3	Осуществляет модифицирование программы МНК-оценивания под реализацию алгоритма включения (исключения). Проводит вычисления и выбирает оптимальную модель для аппроксимации заданных экспериментальных данных, опираясь на полученные значения критериев качества моделей. Оформляет отчет, включающий постановочную часть, текст программы, листинг вычислений
7. Структурное моделирование с использованием внешних критериев селекции моделей	0	2	3	Изучает функциональные возможности программного комплекса ОДА, пользуясь данными методическими рекомендациями и встроенной контекстной помощью. Осуществляет подготовку исходных данных из лабораторных работ № 2, № 6. Решает задачу выбора "наилучшей" модели регрессии с использованием внешних критериев качества. Оформляет отчет, включающий постановочную часть, текст программы, листинг вычислений

4. Самостоятельная работа обучающегося

№	Виды самостоятельной работы	Ссылки на результаты обучения	Часы на выполнение	Часы на консультации
Семестр: 7				
1	РГЗ	3	10	2
Подготовка и выполнение расчетно-графического задания: Статистические методы анализа данных : методические указания к выполнению лабораторных работ и расчетно-графического задания для 4 курса ФПМИ направлений подготовки 01.03.02 - прикладная математика и информатика, 02.03.03 - математическое обеспечение и администрирование информационных систем / Новосиб. гос. техн. ун-т ; [сост. А. А. Попов]. - Новосибирск, 2017. - 33, [2] с.. - Режим доступа: http://elibrary.nstu.ru/source?bib_id=vtls000235022				
2	Подготовка к занятиям	1	40	4
Подготовка и изучение материала по темам лабораторных работ: Статистические методы анализа данных : методические указания к выполнению лабораторных работ и расчетно-графического задания для 4 курса ФПМИ направлений подготовки 01.03.02 - прикладная математика и информатика, 02.03.03 - математическое обеспечение и администрирование информационных систем / Новосиб. гос. техн. ун-т ; [сост. А. А. Попов]. - Новосибирск, 2017. - 33, [2] с.. - Режим доступа: http://elibrary.nstu.ru/source?bib_id=vtls000235022 Статистические методы анализа данных : методические указания для 4 курса ФПМИ всех направлений и специальностей / Новосиб. гос. техн. ун-т ; [сост.: А. А. Попов]. - Новосибирск, 2004. - 31 с.. - Режим доступа: http://elibrary.nstu.ru/source?bib_id=vtls000031994				
3	Подготовка к аттестации	1, 2	12	2

Подготовка к аттестации на основе повторения материала по дисциплине: Организация самостоятельной работы студентов Новосибирского государственного технического университета : методическое руководство / Новосиб. гос. техн. ун-т ; [сост.: Ю. В. Никитин, Т. Ю. Сурнина]. - Новосибирск, 2016. - 19, [1] с. : табл. - Режим доступа: http://elibrary.nstu.ru/source?bib_id=vtls000234042

5. Технология обучения

Для организации и контроля самостоятельной работы обучающихся, а также проведения консультаций применяются информационно-коммуникационные технологии (табл. 5.1).

Таблица 5.1

Деятельность	Информационно-коммуникационные технологии
Информирование	e-mail; Личный типовой сайт
Консультирование	e-mail; Личный типовой сайт
Контроль	e-mail
Размещение учебных материалов	Личный типовой сайт

Таблица 5.2

Активные и интерактивные формы проведения занятий

№	Наименование активных форм
1	Проблемный метод
Краткое описание применения: Обсуждение постановки задачи, вопросов реализации алгоритмов ее решения, полученных результатов	

6. Правила аттестации обучающихся по учебной дисциплине

Для аттестации обучающихся по дисциплине используется балльно-рейтинговая система (БРС), позволяющая выставять оценки по традиционной шкале и 15-уровневой ECTS.

Краткая информация о БРС приведена в табл. 6.1.

Таблица 6.1

Оцениваемые виды деятельности обучающихся	Мин. балл	Максимальный балл
Семестр: 7		
Лабораторная №1:	6	10
Лабораторная №2:	6	10
Лабораторная №3:	6	10
Лабораторная №4:	6	10
Лабораторная №5:	6	10
Лабораторная №6:	6	10
Лабораторная №7:	6	8
РГЗ:	6	12
Зачет:	7	20

В таблице 6.2 представлено соответствие форм контроля заявляемым требованиям к результатам освоения дисциплины.

Таблица 6.2

Коды компетенций ФГОС	Результаты обучения	Формы контроля	
		Защита РГЗ	Зачет
ОПК.1	35. знать методы теории вероятностей и математической статистики		+
ПК.2	34. знать теоретические основы методов построения зависимостей по экспериментальным данным		+
	у2. уметь применять основные математические методы при построении моделей	+	

Фонд оценочных средств по дисциплине представлен в приложении № 1 к рабочей программе.

7. Литература

Основная литература

1. Тимофеев В. С. Эконометрика : [учебник] / В. С. Тимофеев, А. В. Фаддеенков, В. Ю. Щеколдин ; Новосиб. гос. техн. ун-т. - Новосибирск, 2011. - 345 с. : ил., табл., портр.
2. Сидняев Н. И. Теория планирования эксперимента и анализ статистических данных : учебное пособие [для вузов по специальности "Прикладная математика"] / Н. И. Сидняев. - М., 2011. - 399 с. : ил., табл., схемы
3. Малов С. В. Регрессионный анализ. Теоретические основы и практические рекомендации / С. В. Малов ; С.-Петерб. гос. ун-т. - Санкт-Петербург, 2013. - 275 с.
4. Лисицин Д. В. Методы построения регрессионных моделей : [учебное пособие] / Д. В. Лисицин ; Новосиб. гос. техн. ун-т. - Новосибирск, 2011. - 74, [2] с. : табл., ил. - Режим доступа: http://elibrary.nstu.ru/source?bib_id=vtls000154176
5. Бериков В. Б. Эконометрика : учебное пособие / В. Б. Бериков ; Новосиб. гос. техн. ун-т, Фак. приклад. мат. и информатики. - Новосибирск, 2010. - 75, [1] с. : ил., табл. - Режим доступа: http://elibrary.nstu.ru/source?bib_id=vtls000148038
6. Компьютерные технологии анализа данных в эконометрике / Д.М. Дайитбегов. - 2-е изд., испр. и доп. - М.: Вузовский учебник: ИНФРА-М, 2010. - 578 с.: 70x100 1/16. - (Научная книга). (переплет) ISBN 978-5-9558-0191-9 - Режим доступа: <http://znanium.com/catalog.php?bookinfo=251791> - Загл. с экрана.

Дополнительная литература

1. Дрейпер Н. Р. Прикладной регрессионный анализ : пер. с англ. / Норман Р. Дрейпер, Гарри Смит. - М. [и др.], 2007. - 911 с. : ил., табл.
2. Кобзарь А. И. Прикладная математическая статистика : для инженеров и научных работников / А. И. Кобзарь. - М., 2006. - 813 с. : табл.
3. Попов А. А. Конструирование линейных регрессионных моделей с разнотипными переменными : Учеб. пособие для IY курса ФПМИ дневной формы обучения / А. А. Попов; Новосиб. гос. техн. ун-т. - Новосибирск, 1999. - 55 с. : ил.
4. Чубукова И. А. Data Mining : учебное пособие / И. А. Чубукова. - М., 2006. - 382 с. : ил.
5. Воскобойников Ю. Е. Регрессионный анализ данных в пакете Mathcad : учебное пособие / Ю. Е. Воскобойников. - Санкт-Петербург [и др.], 2011. - 223, [1] с. : ил., табл. + 1 CD.

Интернет-ресурсы

1. ЭБС НГТУ : <http://elibrary.nstu.ru/>
2. ЭБС «Издательство Лань» : <https://e.lanbook.com/>

3. ЭБС IPRbooks : <http://www.iprbookshop.ru/>

4. ЭБС "Znaniium.com" : <http://znaniium.com/>

5. :

8. Методическое и программное обеспечение

8.1 Методическое обеспечение

1. Статистические методы анализа данных : методические указания для 4 курса ФПМИ всех направлений и специальностей / Новосиб. гос. техн. ун-т ; [сост.: А. А. Попов]. - Новосибирск, 2004. - 31 с. - Режим доступа: http://elibrary.nstu.ru/source?bib_id=vtls000031994
2. Попов А. А. Статистические методы анализа данных: методические указания к расчетно-графическому заданию для студентов IV курса ФПМИ всех направлений и специальностей [Электронный ресурс] : учебно-методическое пособие / А. А. Попов ; Новосиб. гос. техн. ун-т. - Новосибирск, [2011]. - Режим доступа: http://elibrary.nstu.ru/source?bib_id=vtls000162616. - Загл. с экрана.
3. Организация самостоятельной работы студентов Новосибирского государственного технического университета : методическое руководство / Новосиб. гос. техн. ун-т ; [сост.: Ю. В. Никитин, Т. Ю. Сурнина]. - Новосибирск, 2016. - 19, [1] с. : табл.. - Режим доступа: http://elibrary.nstu.ru/source?bib_id=vtls000234042
4. Статистические методы анализа данных : методические указания к выполнению лабораторных работ и расчетно-графического задания для 4 курса ФПМИ направлений подготовки 01.03.02 - прикладная математика и информатика, 02.03.03 - математическое обеспечение и администрирование информационных систем / Новосиб. гос. техн. ун-т ; [сост. А. А. Попов]. - Новосибирск, 2017. - 33, [2] с. - Режим доступа: http://elibrary.nstu.ru/source?bib_id=vtls000235022

8.2 Специализированное программное обеспечение

1 Python

9. Материально-техническое обеспечение

Специальное оборудование

№	Наименование	Назначение
1	Комплект оборудования мультимедийных аудиторий №2, I-426	Для проведения лекций

«Новосибирский государственный технический университет»

Кафедра прикладной математики
Кафедра теоретической и прикладной информатики

“УТВЕРЖДАЮ”
ДЕКАН ФПМИ
д.т.н., доцент В.С. Тимофеев
“ ” г.

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

УЧЕБНОЙ ДИСЦИПЛИНЫ

Статистические методы анализа данных

Образовательная программа: 01.03.02 Прикладная математика и информатика, профиль: Компьютерное моделирование и информационные технологии

1. Обобщенная структура фонда оценочных средств учебной дисциплины

Обобщенная структура фонда оценочных средств по дисциплине Статистические методы анализа данных приведена в Таблице.

Таблица

Формируемые компетенции	Показатели сформированности компетенций (знания, умения, навыки)	Темы	Этапы оценки компетенций	
			Мероприятия текущего контроля (курсовой проект, РГЗ(Р) и др.)	Промежуточная аттестация (экзамен, зачет)
ОПК.1 способность использовать базовые знания естественных наук, математики и информатики, основные факты, концепции, принципы теорий, связанных с прикладной математикой и информатикой	з5. знать методы теории вероятностей и математической статистики	Задачи статистического анализа данных. Многомерный регрессионный анализ. Интервальное оценивание и проверка гипотез. Многомерный регрессионный анализ. Оценивание параметров Расширения общей линейной модели		Зачет, вопросы 1-109, задача
ПК.2/НИ способность понимать, совершенствовать и применять современный математический аппарат	з4. знать теоретические основы методов построения зависимостей по экспериментальным данным	Выбор общего вида функции регрессии Задачи статистического анализа данных. Многомерный регрессионный анализ. Интервальное оценивание и проверка гипотез. Многомерный регрессионный анализ. Оценивание параметров Расширения общей линейной модели		Зачет, вопросы, 1-109, задача
ПК.2/НИ	у2. уметь применять основные математические методы при построении моделей	Интервальное оценивание неизвестных параметров регрессионных, моделей проверка гипотез Моделирование регрессионных процессов Оценивание неизвестных параметров линейных регрессионных моделей Оценивание неизвестных параметров регрессионных моделей в условиях гетероскедастичности возмущений Оценивание неизвестных параметров регрессионных моделей в условиях мультиколлинеарности Последовательные алгоритмы поиска модели объекта Структурное моделирование с использованием внешних критериев селекции моделей	РГЗ: разделы 1 -8	

2. Методика оценки этапов формирования компетенций в рамках дисциплины.

Промежуточная аттестация по дисциплине проводится в 7 семестре - в форме дифференцированного зачета, который направлен на оценку сформированности компетенций ОПК.1, ПК.2/НИ. Время подготовки на зачете составляет 1,5 часа.

Сформированность компетенций проверяется также при проведении мероприятий текущего контроля, указанных в таблице раздела 1.

В 7 семестре обязательным этапом текущей аттестации является расчетно-графическое задание (работа) (РГЗ(Р)). Требования к выполнению РГЗ(Р), состав и правила оценки сформулированы в паспорте РГЗ(Р).

Общие правила выставления оценки по дисциплине определяются балльно-рейтинговой системой, приведенной в рабочей программе дисциплины.

На основании приведенных далее критериев можно сделать общий вывод о сформированности компетенций ОПК.1, ПК.2/НИ за которые отвечает дисциплина, на разных уровнях.

Общая характеристика уровней освоения компетенций.

Ниже порогового. Уровень выполнения работ не отвечает большинству основных требований, теоретическое содержание курса освоено частично, пробелы могут носить существенный характер, необходимые практические навыки работы с освоенным материалом сформированы не достаточно, большинство предусмотренных программой обучения учебных заданий не выполнены или выполнены с существенными ошибками.

Пороговый. Уровень выполнения работ отвечает большинству основных требований, теоретическое содержание курса освоено частично, но пробелы не носят существенного характера, необходимые практические навыки работы с освоенным материалом в основном сформированы, большинство предусмотренных программой обучения учебных заданий выполнено, некоторые виды заданий выполнены с ошибками.

Базовый. Уровень выполнения работ отвечает всем основным требованиям, теоретическое содержание курса освоено полностью, без пробелов, некоторые практические навыки работы с освоенным материалом сформированы недостаточно, все предусмотренные программой обучения учебные задания выполнены, качество выполнения ни одного из них не оценено минимальным числом баллов, некоторые из выполненных заданий, возможно, содержат ошибки.

Продвинутый. Уровень выполнения работ отвечает всем требованиям, теоретическое содержание курса освоено полностью, без пробелов, необходимые практические навыки работы с освоенным материалом сформированы, все предусмотренные программой обучения учебные задания выполнены, качество их выполнения оценено числом баллов, близким к максимальному.

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Новосибирский государственный технический университет»

Кафедра прикладной математики

Кафедра теоретической и прикладной информатики

Паспорт зачета

по дисциплине «Статистические методы анализа данных», 7 семестр

1. Методика оценки

Зачет проводится в письменной форме, по билетам. В ходе зачета преподаватель вправе задавать студенту дополнительные вопросы из общего перечня (п. 4).

Форма билета для зачета

Дисциплина Статистические методы анализа данных

(наименование дисциплины)

БИЛЕТ №

№	Вопрос	Балло в
1	Перечислите известные Вам методы оценивания параметров линейных регрессионных моделей	1
2	Для чего в теореме Гаусса-Маркова используется предположение A3: $rg(X) = m$.	1
3	Назовите недостаток максимальной сопряженности как меры измерения мультиколлинеарности	1
4	Какова история происхождения термина "ридж-оценки".	1
5	Что означает термин "ошибка спецификации модели".	1
6	Как вводится так называемый "критерий регулярности".	1
7	Покажите, что в случае "недобора" регрессоров соответствующие компоненты вектора параметров оцениваются смещено	1
8	Для чего в теореме Гаусса-Маркова используется предположение A1: $E(\varepsilon) = 0$.	1
9	По каким причинам и на каких этапах исследования исследователь вынужден возвращаться на более ранние этапы исследования зависимости.	1
10	Покажите, что если P - матрица проектирования на Ω ; где $\Omega = R[X]$ - пространство образов, порождаемое столбцами матрицы X , то $P^2 = P$.	1
11	Получите выражение для оценки неизвестной дисперсии σ^2 по параллельным наблюдениям	1
12	Покажите, что в модели, содержащей свободный член, сумма остатков $e_i = y_i - \hat{y}_i$ равна нулю.	1
13	Как можно найти (оценить) неизвестный параметр (коэффициент) автокорреляции ошибок наблюдений.	1
14	Какое наблюдение можно принять за "точку разбалансировки".	1
15	Задача	6
	Итого	20

Задача

Дана таблица наблюдений:

№	x	y
1	-1	-1.1
2	-1	-0.9
3	0	2.1
4	0	1.9
5	1	5.1
6	1	4.9

1. По 6 наблюдениям оценить параметры трех моделей $\eta_1 = \theta_1 + \theta_2 x$, $\eta_2 = \theta_1 + \theta_2 x^2$, $\eta_3 = \theta_1 x + \theta_2 x^2$. Какая из этих моделей лучшая по критерию Мэлоуса? Для вычисления критерия Мэлоуса принять, что дисперсия помехи равна $\sigma^2 = 0.01$.

2. Для модели $\eta_1 = \theta_1 + \theta_2 x$ проверить гипотезу $H : \theta_2 = 3$. Квантиль $F_{\alpha,1,4} = 7.71$. Принимается или отклоняется эта гипотеза?

Составитель д.т.н., проф. Попов А.А.

_____ (подпись)

« ____ » _____ 20 ____ г.

Заведующий кафедрой Д.т.н. В.М. Чубич

_____ (подпись)

« ____ » _____ 20 ____ г.

2. Критерии оценки

- Ответ считается **неудовлетворительным**, если оценка составляет от 0 до 6 баллов, Студент при ответе на вопросы не дает определений основных понятий, при решении задачи допускает принципиальные ошибки;

- Ответ засчитывается на **пороговом** уровне, если оценка составляет от 7 до 10 баллов. Студент при ответе на вопросы дает определение основных понятий, знает базовые алгоритмы, при решении задачи допускает непринципиальные ошибки;
- Ответ засчитывается на **базовом** уровне, если оценка составляет от 11 до 15 баллов. Студент при ответе на вопросы формулирует основные понятия, методы, алгоритмы, проводит анализ причин, условий, не допускает ошибок при решении задачи;
- Ответ засчитывается на **продвинутом** уровне, если оценка составляет от 16 до 20 баллов. Студент владеет математическим аппаратом, методами прикладной статистики, демонстрирует глубокие знания и навыки решения задачи построения зависимости.

3. Шкала оценки

Итоговая аттестация по курсу проводится в виде зачета. Для получения допуска к зачету студент должен, как правило, выполнить и защитить все лабораторные работы, выполнить РГЗ. Основным критерий допуска к зачету – число набранных баллов в течение семестра, которое должно составить не менее 43 баллов.

Максимальное количество набираемых баллов в период итоговой аттестации равно 20 баллам. Если студент по результатам итоговой аттестации набирает **менее 7 баллов**, то ему выставляется оценка "неудовлетворительно" уровня FX вне зависимости от числа набранных баллов в семестре с возможностью пересдачи. При успешной пересдаче ему выставляется оценка "удовлетворительно" уровня E. Общее количество набранных баллов по дисциплине определяется простым суммированием набранных баллов по лабораторным работам, РГЗ и на зачете.

Перевод набранного количества баллов в 15 бальную систему и 4-х бальную систему осуществляется по следующей схеме:

ECTS	A+	A	A-	B+	B	B-	C+	C	C-	D+	D	D-	E	FX	F
Баллы	96-100	93-95	90-92	87-89	83-86	80-82	76-79	73-75	70-72	66-69	63-65	60-62	50-59	25-49	0-24
	отлично				хорошо				удовлетворительно				неудовлетв.		

4. Вопросы к зачету по дисциплине «Статистические методы анализа данных»

Полный перечень вопросов

по дисциплине "Статистические методы анализа данных"

1. Назовите основные типы прикладных целей исследования зависимостей.

2. Перечислите типовые задачи одномерного и многомерного анализа данных.
3. Что понимается под терминами "Data Mining", "Web Mining".
4. Назовите разделы многомерного статистического анализа, которые используются в задачах анализа зависимостей при различных комбинациях типов факторов.
5. Назовите основные этапы статистического исследования зависимостей.
6. Что обозначают термины "пассивный" и "активный" эксперимент.
7. Почему этап построения зависимостей под названием "определение класса допустимых решений" называют также этапом параметризации модели.
8. Что означает понятие "линейная по параметрам регрессионная модель".
9. Какова история происхождения термина "Регрессионный анализ".
10. Перечислите известные Вам методы оценивания параметров линейных регрессионных моделей.
11. Что понимается под термином "структура" модели.
12. Как Вы понимаете термин "адекватность" модели.
13. Что означает термин "линейная оценка".
14. Какими свойствами обладает наилучшая линейная оценка.
15. Назовите предположения, которые используются в теореме Гаусса-Маркова.
16. Для чего в теореме Гаусса-Маркова используется предположение A3: $rg(X) = m$.
17. Для чего в теореме Гаусса-Маркова используется предположение A1: $E(\varepsilon) = 0$.
18. Если ошибки наблюдения распределены по нормальному закону, то по какому закону распределены НЛО оценки параметров линейной регрессионной модели и почему.
19. Что означает термин "интервальное оценивание" параметров модели.
20. Как запишется гипотеза о проверке незначимости параметра.
21. Как запишется гипотеза о проверке незначимости регрессии.
22. Дайте толкование выражения "несферические возмущения".
23. Как можно найти (оценить) неизвестный параметр (коэффициент) автокорреляции ошибок наблюдений.
24. Что означает термин "мультиколлинеарность".
25. Что означает термин "проверка структурных изменений".
26. Как обнаружить аномальное наблюдение, используя "правило трех сигм".
27. Перечислите основные меры измерения эффекта мультиколлинеарности.
28. Какие меры измерения эффекта мультиколлинеарности свободны от влияния масштаба факторов.
29. Обобщенный МНК и ридж-оценивание – это одно и то же?
30. Какова история происхождения термина "ридж-оценки".
31. Чем можно руководствоваться при выборе параметра λ в процедуре ридж-оценивания.
32. Смещенными или несмещенными являются ридж-оценки.
33. Какое свойство ридж-оценок можно считать главным и почему.
34. Как оценить число главных компонент.
35. Для чего ведется анализ геометрической структуры исходных данных.
36. В каких целях используются корреляционные поля. Что это за понятие.
37. Что понимается под термином "полная модель".
38. Что означает термин "ошибка спецификации модели".
39. К каким последствиям приводят те или иные ошибки спецификации модели.
40. Перечислите критерии качества моделей, основанные на использовании одной выборки.
41. Что понимается под термином "внешние критерии" качества моделей.
42. Почему предпочтительнее использовать симметричные внешние критерии нежели несимметричные.
43. Назовите несколько (два – три) внешних критериев, относящихся к группе "критерии точности".
44. Назовите несколько (два – три) внешних критериев, относящихся к группе "критерии согласованности".
45. Как вводится так называемый "критерий регулярности".

46. Что понимается под помехоустойчивостью внешних критериев.
47. Почему абсолютно помехоустойчивый критерий получил такое название.
48. Если наблюдений не очень много, то какой критерий качества моделей целесообразно использовать при выборе модели оптимальной сложности.
49. Опишите в общих чертах многорядный комбинаторный алгоритм.
50. Как сказывается влияние латентных факторов на поведение отклика системы.
51. По каким причинам и на каких этапах исследования исследователь вынужден возвращаться на более ранние этапы исследования зависимости.
52. Назовите причины, по которым отклик системы считается случайной величиной.
53. Покажите, что НЛО оценка параметров линейной регрессионной модели является несмещенной.
54. Покажите на простом примере, что матрица $X^T X$ будет бесконечно возрастать при стремлении числа наблюдений к ∞ .
55. Покажите, что дисперсия НЛО оценки равна $D(\hat{\theta}) = \sigma^2 (X^T X)^{-1}$.
56. Покажите, что если P - матрица проектирования на Ω ; где $\Omega = R[X]$ - пространство образов, порождаемое столбцами матрицы X , то $P^2 = P$.
57. Покажите, что если P - матрица проектирования на Ω ; где $\Omega = R[X]$ - пространство образов, порождаемое столбцами матрицы X , то $(I - P)^2 = I - P$.
58. Покажите, что если P - матрица проектирования на Ω ; где $\Omega = R[X]$ - пространство образов, порождаемое столбцами матрицы, то $(I_n - P)$ - матрица ортогонального проектирования на $\Omega^\perp$.
59. Покажите, что поскольку $y = \hat{y} + \hat{e}$, то $y^T y = \hat{y}^T \hat{y} + \hat{e}^T \hat{e}$, где $\hat{y} = Py$, $\hat{e} = (I - P)y$.
60. Покажите, что $\hat{e}^T X = 0$, где $\hat{e}$ - остатки, X - матрица наблюдений.
61. Получите выражение для оценки неизвестной дисперсии σ^2 по методу максимального правдоподобия.
62. Получите выражение для оценки неизвестной дисперсии σ^2 по параллельным наблюдениям.
63. Как проверить незначимость параметра линейной регрессионной модели.
64. Покажите, что в модели, состоящей только из свободного члена, МНК оценка этого параметра равна среднему значению отклика по выборке.
65. Что понимается под доступным обобщенным методом наименьших квадратов.
66. Получите выражение для ковариационной матрицы ошибок наблюдений, если известно, что они описываются процессом авторегрессии первого порядка.
67. Каким недостатком обладает Тест Дарбина-Уотсона.
68. Покажите, что в модели, содержащей свободный член, сумма остатков $e_i = y_i - \hat{y}_i$ равна нулю.
69. Что позволяет обнаружить проведенный анализ остатков.
70. Какие наблюдения считаются "плохо влияющими".
71. Какое наблюдение можно принять за "точку разбалансировки".
72. Покажите, что при усилении эффекта мультиколлинеарности дисперсия оценок параметров линейной регрессионной модели может бесконечно возрастать.
73. Как освободить некоторые меры измерения эффекта мультиколлинеарности от влияния масштаба факторов.
74. Как узнать какие именно входные факторы ответственны за возникновение эффекта мультиколлинеарности.
75. Назовите недостатки и достоинства максимальной парной сопряженности как меры измерения мультиколлинеарности.
76. Назовите недостаток максимальной сопряженности как меры измерения мультиколлинеарности.
77. Покажите, что в среднем норма оценок параметров регрессионной модели при усилении эффекта мультиколлинеарности (например, при стремлении минимального собственного числа к 0) стремится к бесконечности.

78. Покажите, что ридж-оценка параметров линейной регрессионной модели является линейным преобразованием МНК оценки.
79. Какими свойствами должны обладать так называемые главные компоненты.
80. Объясните почему коэффициент детерминации при добавлении в модель регрессоров ведет себя как возрастающая функция от числа регрессоров.
81. Если для одной модели $R^2 = 0.5$, а для второй $R^2 = 0.75$, то во сколько раз остаточная сумма квадратов для второй модели меньше остаточной суммы квадратов первой модели.
82. Поясните, почему можно считать, что адекватными могут быть признаны многие модели.
83. Покажите, что в случае "недобора" регрессоров соответствующие компоненты вектора параметров оцениваются смещено.
84. Поясните, почему следует ожидать, что графики зависимостей внешних критериев качества моделей от числа включенных в модель регрессоров, будут иметь минимум.
85. Дайте интерпретацию гипотезы о структурных изменениях с позиции использования внешних критериев.
86. Как вводится субидеальный критерий стабильности на основе повторных наблюдений.
87. Назовите основные свойства оператора выметания. Какова оптимальная последовательность выметаний.
88. Что представляют собой алгоритмы включения и исключения.
89. Эффективный алгоритм построения всех регрессий. Дать описание алгоритма.
90. Многорядный комбинаторный алгоритм. Дать описание алгоритма.
91. Внешние критерии качества модели. Обосновать необходимость их использования. Классификация. Ввести и описать 6 критериев.
92. Внутренние критерии качества модели. Вывести критерий Мэллоуса.
93. Выбор наилучшей модели. К каким последствиям приводят ошибки спецификации модели типа "недобор" или "перебор" регрессоров. Показать выкладками.
94. Выбор общего вида функции регрессии (структуры модели). Использование априорной информации, предварительный анализ геометрической структуры исходных данных, учет требований гладкости.
95. Гетероскедастичность возмущений. Проверка данных на гетероскедастичность (тест Гольдфельда-Квандтона).
96. Гетероскедастичность возмущений. Проверка данных на гетероскедастичность (обобщение теста Гольдфельда-Квандтона).
97. Наилучшие линейные оценки (НЛЮ). Теорема Гаусса-Маркова (с доказательством).
98. Обобщенный МНК.
99. Оценивание неизвестной дисперсии ошибки наблюдения.
100. Оценивание параметров в модели регрессии при наличии линейных ограничений.
101. Оценивание параметров в условиях мультиколлинеарности. Меры измерения мультиколлинеарности. Как определить какие факторы ответственны за возникновение мультиколлинеарности.
102. Оценивание параметров в условиях повторных наблюдений.
103. Оценивание параметров регрессионной модели по методу максимального правдоподобия (ММП).
104. Помехоустойчивость внешних критериев.
105. Проверка гипотез и внешние критерии качества модели.
106. Проверка структурных изменений. Формулирование гипотезы и ее проверка.
107. Ридж-оценки неизвестных параметров и их свойства.
108. Свойства МНК оценки параметров линейной регрессионной модели.
109. Совместное и раздельное интервальное оценивание параметров линейных регрессионных моделей.

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Новосибирский государственный технический университет»

Кафедра теоретической и прикладной информатики

**Паспорт
расчетно-графического задания (работы)**

по дисциплине «Статистические методы анализа данных», 7 семестр

1. Методика оценки

При выполнении расчетно-графического задания (работы) студенты должны провести полный цикл исследований по построению регрессионной зависимости по имеющимся экспериментальным данным.

В перечень исследований как обязательные части должны входить:

1. проверка данных на мультиколлинеарность;
2. проверка данных на гетероскедастичность (предположительно, что чем дальше от центра эксперимента проведено наблюдение, то возможно дисперсия его больше);
3. проверка данных на автокорреляцию (упорядоченность наблюдений по своим номерам считать упорядоченностью по времени);
4. выбор предварительного состава регрессоров с использованием корреляционных полей. В качестве регрессоров-кандидатов предположительно могут выступать: свободный член, сами факторы, их взаимодействия (двух-трех факторов), квадраты факторов;
5. выбор модели оптимальной сложности с использованием критериев Мэллоуса, скорректированного коэффициента детерминации, внешних критериев;
6. проверка адекватности выбранной модели с использованием повторных наблюдений (последние 6 наблюдений выборки), по которым необходимо будет вычислить оценку дисперсии наблюдений;
7. построение графиков остатков в различных координатах (по номеру наблюдений, по факторам, по отклику);
8. определение, опираясь на построенную модель, точки в факторном пространстве, имеющей максимальное значение математического ожидания отклика. Вычисление для этой точки доверительного интервала. Координаты такой точки не обязательно должны совпадать с какой-либо точкой из имеющихся в таблице наблюдений.

Оцениваемые позиции: части 1-8.

2. Критерии оценки

- Работа считается **не выполненной**, если большинство частей работ не выполнены, оценка составляет 1-2 баллов.
- Работа считается выполненной **на пороговом** уровне, если части РГЗ(Р) выполнены

формально: не проведен анализ данных на гетероскедастичность, автокорреляцию, мультиколлинеарность. Оценка составляет 3-5 _____ баллов.

- Работа считается выполненной **на базовом** уровне, если анализ задачи выполнен в полном объеме, однако структура модели подобрана неверно, оценка составляет 6-9 _____ баллов.
- Работа считается выполненной **на продвинутом** уровне, если анализ данных выполнен в полном объеме, структура модели подобрана верно, оценка составляет 10-12 _____ баллов.

3. Шкала оценки

В общей оценке по дисциплине баллы за РГЗ(Р) учитываются в соответствии с правилами балльно-рейтинговой системы, приведенными в рабочей программе дисциплины.

4. Примерный перечень тем РГЗ(Р)

РГЗ по дисциплине является типовым. Перечень разделов определен в п.1. Экспериментальные данные представляются в виде таблицы наблюдений типа "вход-выход" в формате *xlsx*. Номер варианта задания (v.1, v.2,...) соответствует порядковому номеру студента в списке группы.

Общие правила выставления оценки по дисциплине определяются балльно-рейтинговой системой, приведенной в рабочей программе дисциплины.

На основании приведенных далее критериев можно сделать общий вывод о сформированности компетенций ОПК.1, ОПК.2, ПК.3/ПК, ПК.8.В/ПК, за которые отвечает дисциплина, на разных уровнях.