

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

Кафедра теоретической и прикладной информатики

“УТВЕРЖДАЮ”
ДЕКАН ФПМИ

д.т.н. Тимофеев В. С.

“ ____ ” _____ г.

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ
Статистические методы анализа данных

Образовательная программа: 02.03.03 Математическое обеспечение и администрирование информационных систем, профиль: Математическое и программное обеспечение информационных технологий

Курс: 4, семестр : 7

Факультет прикладной математики и информатики,

		Семестр
№	Вид деятельности	7
1	Всего зачетных единиц (кредитов)	5
2	Всего часов	180
3	Всего занятий в контактной форме, час.	50
4	Лекции, час.	18
5	Практические занятия, час.	0
6	Лабораторные занятия, час.	18
7	из них в активной и интерактивной форме, час.	6
8	Аттестация, час.	2
9	Консультации, час.	12
10	Самостоятельная работа, час.	130
11	Виды самостоятельной работы (курсовой проект, курсовая работа, РГЗ, подготовка к контрольной работе)	РГЗ
12	Вид аттестации	Э

Рабочая программа составлена на основании федерального государственного образовательного стандарта высшего образования по направлению (специальности): 02.03.03 Математическое обеспечение и администрирование информационных систем

ФГОС введен в действие приказом №222 от 12.03.2015 г. , дата утверждения: 07.04.2015 г.

Место дисциплины в структуре учебного плана: Блок 1, базовая

Рабочая программа разработана на основе компетентностной модели выпускника по направлению (специальности): 02.03.03 Математическое обеспечение и администрирование информационных систем

Рабочая программа обсуждена на заседании кафедры

ТПИ, протокол заседания кафедры №4 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.

Ответственный за образовательную программу:

заведующий кафедрой Чубич В. М.

1. Внешние требования

Таблица 1.1

Компетенция ФГОС: ОПК.1 способность решать стандартные задачи профессиональной деятельности на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий и с учетом основных требований информационной безопасности; в части следующих результатов обучения:	
у5. Уметь использовать специализированные программные средства при решении профессиональных задач	
Компетенция ФГОС: ОПК.2 способность применять в профессиональной деятельности знания математических основ информатики; в части следующих результатов обучения:	
з8. Знать теоретические основы методов построения зависимостей по экспериментальным данным	
у6. Уметь использовать базовые знания алгебры, вычислительной математики, теории вероятностей при разработке вычислительных алгоритмов построения многофакторных моделей зависимостей по экспериментальным данным	
Компетенция ФГОС: ПК.3 готовность к разработке моделирующих алгоритмов и реализации их на базе языков и пакетов прикладных программ моделирования; в части следующих результатов обучения:	
у4. Уметь реализовывать вычислительные алгоритмы построения зависимостей по экспериментальным данным	
Компетенция НГТУ: ПК.8.В/ПК готовность к использованию основных моделей информационных технологий для решения прикладных задач; в части следующих результатов обучения:	
у3. Владеть методами построения и анализа линейных моделей по экспериментальным данным	

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.у5 уметь использовать специализированные программные средства при решении профессиональных задач	
1.уметь использовать специализированные программные средства при решении профессиональных задач	Лабораторные работы; Самостоятельная работа
ОПК.2.з8 Знать теоретические основы методов построения зависимостей по экспериментальным данным	
2.Знать теоретические основы методов построения зависимостей по экспериментальным данным	Лекции; Самостоятельная работа
ОПК.2.у6 Уметь использовать базовые знания алгебры, вычислительной математики, теории вероятностей при разработке вычислительных алгоритмов построения многофакторных моделей зависимостей по экспериментальным данным	
3.Уметь использовать базовые знания алгебры, вычислительной математики, теории вероятностей при разработке вычислительных алгоритмов построения многофакторных моделей зависимостей по экспериментальным данным	Лекции; Самостоятельная работа
ПК.8.В/ПК.у3 Владеть методами построения и анализа линейных моделей по экспериментальным данным	
4.Владеть методами построения и анализа линейных моделей по экспериментальным данным	Лабораторные работы
ПК.3.у4 Уметь реализовывать вычислительные алгоритмы построения зависимостей по экспериментальным данным	
5.Уметь реализовывать вычислительные алгоритмы построения зависимостей по экспериментальным данным	Лабораторные работы

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.				
1. Задачи статистического исследования зависимостей, задачи снижения размерности, задачи классификации, задачи кластеризации. Методы интеллектуального анализа данных (Data Mining)	0	2	2	Учебная деятельность состоит в изучении теоретического материала, подготовки и выполнения лабораторных работ, их защита
Дидактическая единица: Параметрические регрессионные схемы.				
2. Основные требования, предъявляемые к оценкам параметров. Наилучшие линейные оценки (НЛО). Вычисление НЛО, их свойства. Связь НЛО с оценками по методу наименьших квадратов (МНК).	0	4	2, 3	Учебная деятельность состоит в изучении теоретического материала, подготовки и выполнения лабораторных работ, их защита
Дидактическая единица: Интервальное оценивание и проверка гипотез				
3. Интервальное оценивание и проверка гипотез. Оценивание неизвестной дисперсии ошибки наблюдения. Проверка адекватности регрессионной модели.	0	2	2, 3	Учебная деятельность состоит в изучении теоретического материала, подготовки и выполнения лабораторных работ, их защита
4. Оценивание параметров в модели регрессии при наличии линейных ограничений. Проверка общих линейных гипотез	0	2	2, 3	Учебная деятельность состоит в изучении теоретического материала, подготовки и выполнения лабораторных работ, их защита
Дидактическая единица: Расширения линейной модели				
5. Оценивание параметров в условиях гетероскедастичности, автокорреляции ошибок наблюдения, в условиях мультиколлинеарности	0	4	2	Учебная деятельность состоит в изучении теоретического материала, подготовки и выполнения лабораторных работ, их защита
Дидактическая единица: Выбор общего вида функции регрессии				
6. Критерии качества модели. Алгоритмы выбора наилучшей модели регрессии: алгоритмы включения и исключения, эффективный алгоритм построения всех регрессий, многорядный комбинаторный алгоритм	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.				

1. Моделирование регрессионных процессов	1	2	4	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	2	1, 4, 5	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				
3. Интервальное оценивание неизвестных параметров регрессионных, моделей проверка гипотез	1	4	1, 4, 5	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: Расширения линейной модели				

4. Оценивание неизвестных параметров регрессионных моделей в условиях гетероскедастичности возмущений	1	2	1, 4, 5	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	1, 4, 5	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. Сравняет 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				

6. Последовательные алгоритмы поиска модели объекта	1	4	1, 4, 5	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	1, 4, 5	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	1, 3	30	4
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	2	70	6
Подготовка и изучения материалов по теме 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	2, 3	30	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	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		
<i>Лабораторная №1:</i>	3	7
Контролирующие материалы - список вопросов		
<i>Лабораторная №2:</i>	3	7
Контролирующие материалы - список вопросов		
<i>Лабораторная №3:</i>	3	7
Контролирующие материалы - список вопросов		
<i>Лабораторная №4:</i>	3	7
Контролирующие материалы - список вопросов		
<i>Лабораторная №5:</i>	3	7
Контролирующие материалы - список вопросов		
<i>Лабораторная №6:</i>	3	7

Контролирующие материалы - список вопросов		
Лабораторная №7:	3	7
Контролирующие материалы - список вопросов		
Лабораторная №8:	3	6
Контролирующие материалы - список вопросов		
РГЗ:	3	12
Контролирующие материалы - Защита выбранных вариантов решений		
Экзамен:	14	40
Контролирующие материалы - список вопросов		

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	Экзамен
ОПК.1	у5. уметь использовать специализированные программные средства при решении профессиональных задач	+	
ОПК.2	з8. Знать теоретические основы методов построения зависимостей по экспериментальным данным		+
	у6. Уметь использовать базовые знания алгебры, вычислительной математики, теории вероятностей при разработке вычислительных алгоритмов построения многофакторных моделей зависимостей по экспериментальным данным		+
ПК.3	у4. Уметь реализовывать вычислительные алгоритмы построения зависимостей по экспериментальным данным	+	+
	ПК.8.В/ПК у3. Владеть методами построения и анализа линейных моделей по экспериментальным данным	+	+

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. 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
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. Малов С. В. Регрессионный анализ. Теоретические основы и практические рекомендации / С. В. Малов ; С.-Петерб. гос. ун-т. - Санкт-Петербург, 2013. - 275 с.

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 IV 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 "Znanium.com" : <http://znanium.com/>
5. :

8. Методическое и программное обеспечение

8.1 Методическое обеспечение

1. 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.
2. 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. 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

Презентационное оборудование

№	Наименование	Назначение
1	Презентационное оборудование (мультимедиа-проектор, экран, компьютер для управления)	Для проведения лекций

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

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: 02.03.03 Математическое обеспечение и администрирование информационных систем, профиль: Математическое и программное обеспечение информационных технологий

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 решать стандартные задачи профессиональной деятельности на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий и с учетом основных требований информационной безопасности	у5. уметь использовать специализированные программные средства при решении профессиональных задач	Интервальное оценивание неизвестных параметров регрессионных, моделей проверка гипотез 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	Экзамен, вопросы 1-109, задача
ОПК.2 способность применять в профессиональной деятельности знания математических основ информатики	з8. Знать теоретические основы методов построения зависимостей по экспериментальным данным	Критерии качества модели. Алгоритмы выбора наилучшей модели регрессии: алгоритмы включения и исключения, эффективный алгоритм построения всех регрессий, многорядный комбинаторный алгоритм Задачи статистического исследования зависимостей, задачи снижения размерности, задачи классификации, задачи кластеризации. Интервальное оценивание и проверка гипотез. Оценивание неизвестной дисперсии ошибки наблюдения. Проверка адекватности регрессионной модели. Основные требования, предъявляемые к оценкам параметров. Наилучшие линейные оценки (НЛО). Вычисление НЛО, их свойства. Связь НЛО с оценками по методу наименьших квадратов (МНК). Оценивание параметров в модели регрессии при наличии линейных ограничений. Проверка общих линейных гипотез Оценивание		Экзамен. Вопросы 1-12, 14-34, 51-109

		параметров в условиях гетероскедастичности, автокорреляции ошибок наблюдения, в условиях мультиколлинеарности		
ОПК.2	уб. Уметь использовать базовые знания алгебры, вычислительной математики, теории вероятностей при разработке вычислительных алгоритмов построения многофакторных моделей зависимостей по экспериментальным данным	Интервальное оценивание и проверка гипотез. Оценивание неизвестной дисперсии ошибки наблюдения. Проверка адекватности регрессионной модели. Основные требования, предъявляемые к оценкам параметров. Наилучшие линейные оценки (НЛО). Вычисление НЛО, их свойства. Связь НЛО с оценками по методу наименьших квадратов (МНК). Оценивание параметров в модели регрессии при наличии линейных ограничений. Проверка общих линейных гипотез		Экзамен. Вопросы 13-21, 51-109, задача
ПК.3/ПК готовность к разработке моделирующих алгоритмов и их реализации их на базе современных языков и программных систем	у4. Уметь реализовывать вычислительные алгоритмы построения зависимостей по экспериментальным данным	Интервальное оценивание неизвестных параметров регрессионных, моделей проверка гипотез 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	Экзамен, вопросы 1-109, задача
ПК.8.В/ПК готовность к использованию основных моделей информационных технологий для решения прикладных задач	у3. Владеть методами построения и анализа линейных моделей по экспериментальным данным	Интервальное оценивание неизвестных параметров регрессионных, моделей проверка гипотез 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	Экзамен, вопросы 1-109, задача

2. Методика оценки этапов формирования компетенций в рамках дисциплины.

Промежуточная аттестация по дисциплине проводится в 7 семестре - в форме экзамена, который направлен на оценку сформированности компетенций ОПК.1, ОПК.2, ПК.3/ПК, ПК.8.В/ПК. Время подготовки на экзамене 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).

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.

КОМПЛЕКТ
контролирующих материалов
по дисциплине "Статистические методы анализа данных"
для студентов направления 02.03.03 Математическое обеспечение и администрирование
информационных систем, профиль: Математическое и программное обеспечение
информационных технологий

Дисциплина "Статистические методы анализа данных"

Экзаменационный билет № 1

№	Вопрос	Баллов
1	Назовите основные типы прикладных целей исследования зависимостей.	1
2	Почему этап построения зависимостей под названием "определение класса допустимых решений" называют также этапом параметризации модели.	1
3	Что означает термин "линейная оценка".	1
4	Назовите недостатки и достоинства максимальной парной сопряженности как меры измерения мультиколлинеарности	1
5	Перечислите основные меры измерения эффекта мультиколлинеарности	1
6	Как оценить число главных компонент.	2
7	Что понимается под термином "внешние критерии" качества моделей.	2
8	Если наблюдений не очень много, то какой критерий качества моделей целесообразно использовать при выборе модели оптимальной сложности.	2
9	Покажите, что дисперсия НЛО оценки равна $D(\hat{\theta}) = \sigma^2 (X^T X)^{-1}$.	2
10	Получите выражение для оценки неизвестной дисперсии σ^2 по методу максимального правдоподобия	2
11	Проверка данных на автокорреляцию. Тест Дарбина-Уотсона.	2
12	Как запишется гипотеза о проверке незначимости параметра и как ее проверить.	2
13	Какими свойствами должны обладать так называемые главные компоненты.	2
14	Назовите основные свойства оператора выметания. Какова оптимальная последовательность выметаний - так называемый "счетчик Гарсайда".	2
15	Внутренние критерии качества модели. Вывести критерий Мэллоуса	5
16	Задача	12
	Итого	40

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 Мэллоуса? Для вычисления критерия Мэллоуса 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.

Дисциплина "Статистические методы анализа данных"

Билет № 2

№	Вопрос	Баллов
1	Перечислите типовые задачи одномерного и многомерного анализа данных.	1
2	Что означает понятие "линейная по параметрам регрессионная модель".	1
3	Какими свойствами обладает наилучшая линейная оценка.	1
4	Почему предпочтительнее использовать симметричные внешние критерии нежели несимметричные	1
5	Какие меры измерения эффекта мультиколлинеарности свободны от влияния масштаба факторов	1
6	Для чего ведется анализ геометрической структуры исходных данных	2
7	Как запишется гипотеза о проверке незначимости регрессии и как ее проверить	2
8	Что понимается под терминами "Data Mining", "Web Mining". Перечислите основные задачи.	2
9	Покажите на простом примере, что матрица $X^T X$ будет бесконечно возрастать при стремлении числа наблюдений к ∞ .	2
10	Покажите, что $\hat{e}^T X = 0$, где $\hat{e}$ - остатки, X – матрица наблюдений.	2
11	Получите выражение для ковариационной матрицы ошибок наблюдений, если известно, что они описываются процессом авторегрессии первого порядка	2
12	Как узнать какие именно входные факторы ответственны за возникновение эффекта мультиколлинеарности	2
13	Объясните почему коэффициент детерминации при добавлении в модель регрессоров ведет себя как возрастающая функция от числа регрессоров.	2
14	Как вводится субидеальный критерий стабильности на основе повторных наблюдений.	2
15	Наилучшие линейные оценки (НЛО). Теорема Гаусса-Маркова (с доказательством эффективности оценок).	5
16	Задача	12
	Итого	40

Задача

Дана таблица наблюдений:

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

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 $H: \theta_1 = 0$. 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.

Дисциплина "Статистические методы анализа данных"

Билет № 3

№	Вопрос	Баллов
1	Назовите разделы многомерного статистического анализа, которые используются в задачах анализа зависимостей при различных комбинациях типов факторов.	1
2	Какова история происхождения термина "Регрессионный анализ".	1
3	Назовите предположения, которые используются в теореме Гаусса-Маркова.	1
4	Дайте толкование выражения "несферические возмущения".	1
5	Обобщенный МНК и ридж-оценивание – это одно и то же?	1
6	В каких целях используются корреляционные поля. Что это за понятие. Примеры.	2
7	Назовите несколько (два – три) внешних критериев, относящихся к группе "критерии точности".	2
8	Назовите основные этапы статистического исследования зависимостей	2
9	Покажите, что НЛО оценка параметров линейной регрессионной модели является несмещенной	2
10	Покажите, что поскольку $y = \hat{y} + \hat{e}$, то $y^T y = \hat{y}^T \hat{y} + \hat{e}^T \hat{e}$, где $\hat{y} = Py$, $\hat{e} = (I - P)y$.	2
11	Что понимается под доступным обобщенным методом наименьших квадратов	2
12	Как освободить некоторые меры измерения эффекта мультиколлинеарности от влияния масштаба факторов.	2
13	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 для первой модели.	2
14	Дайте интерпретацию гипотезы о структурных изменениях с позиции использование внешних критериев.	2
15	Наилучшие линейные оценки (НЛО). Теорема Гаусса-Маркова (с доказательством состоятельности оценок).	5
16	Задача	12
	Итого	40

Задача

Дана таблица наблюдений:

№	x	y
1	-1	5.1
2	-1	4.9
3	0	3.1
4	0	2.9
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 $\eta_1 = \theta_1 + \theta_2 x^2$ проверить гипотезу $H: \theta_2 = 2$. 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.

Дисциплина "Статистические методы анализа данных"

Билет № 4

№	Вопрос	Баллов
1	Что обозначают термины "пассивный" и "активный" эксперимент.	1
2	Как Вы понимаете термин "адекватность" модели.	1
3	Что означает термин "интервальное оценивание" параметров модели	1
4	Как обнаружить аномальное наблюдение, используя "правило трех сигм".	1
5	Какое свойство ридж-оценок можно считать главным и почему	1
6	Перечислите критерии качества моделей, основанные на использовании одной выборки.	2
7	Почему абсолютно помехоустойчивый критерий получил такое название.	2
8	Опишите в общих чертах многорядный комбинаторный алгоритм	2
9	Покажите на простом примере, что матрица $X^T X$ будет бесконечно возрастать при стремлении числа наблюдений к ∞ .	2
10	Покажите, что если P - матрица проектирования на Ω ; где $\Omega = R[X]$ - пространство образов, порожаемое столбцами матрицы, то $(I_n - P)$ - матрица ортогонального проектирования на $\Omega^\perp$.	2
11	Покажите, что в модели, состоящей только из свободного члена, МНК оценка этого параметра равна среднему значению отклика по выборке	2
12	Покажите, что при усилении эффекта мультиколлинеарности дисперсия оценок параметров линейной регрессионной модели может бесконечно возрастать	2
13	Покажите, что ридж-оценка параметров линейной регрессионной модели является линейным преобразованием МНК оценки	2
14	Поясните, почему следует ожидать, что графики зависимостей внешних критериев качества моделей от числа включенных в модель регрессоров, будут иметь минимум.	2
15	Несмещенное оценивание неизвестной дисперсии ошибки наблюдения	5
16	Задача	12
	Итого	40

Задача

Дана таблица наблюдений:

№	x	y
1	-1	1.1
2	-1	0.9
3	0	3.1
4	0	2.9
5	1	1.1
6	1	0.9

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 $\eta_1 = \theta_1 + \theta_2 x^2$ проверить гипотезу $H: \theta_1 = 0$. 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.

Дисциплина "Статистические методы анализа данных"

Билет № 5

№	Вопрос	Баллов
1	Перечислите известные Вам методы оценивания параметров линейных регрессионных моделей	1
2	Для чего в теореме Гаусса-Маркова используется предположение А3: $rg(X) = m$.	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".	2
7	Покажите, что в случае "недобора" регрессоров соответствующие компоненты вектора параметров оцениваются смещено	2
8	Для чего в теореме Гаусса-Маркова используется предположение А1: $E(\varepsilon) = 0$.	2
9	По каким причинам и на каких этапах исследования исследователь вынужден возвращаться на более ранние этапы исследования зависимости.	2
10	Покажите, что если P - матрица проектирования на Ω ; где $\Omega = R[X]$ - пространство образов, порожаемое столбцами матрицы X , то $P^2 = P$.	2
11	Получите выражение для оценки неизвестной дисперсии σ^2 по параллельным наблюдениям	2
12	Покажите, что в модели, содержащей свободный член, сумма остатков $e_i = y_i - \hat{y}_i$ равна нулю.	2
13	Как можно найти (оценить) неизвестный параметр (коэффициент) автокорреляции ошибок наблюдений.	2
14	Какое наблюдение можно принять за "точку разбалансировки".	2
15	Оценивание параметров в модели регрессии при наличии линейных ограничений	5
16	Задача	12
	Итого	40

Задача

Дана таблица наблюдений:

№	x	y
1	-1	-1.1
2	-1	-0.9
3	0	0.1
4	0	-0.1
5	1	3.1
6	1	2.9

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 $\eta_1 = \theta_1 + \theta_2 x^2$ проверить гипотезу $H: \theta_2 = 0$. 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.

Дисциплина "Статистические методы анализа данных"

Билет № 6

№	Вопрос	Баллов
1	Что понимается под термином "структура" модели.	1
2	Если ошибки наблюдения распределены по нормальному закону, то по какому закону распределены НЛО оценки параметров линейной регрессионной модели и почему.	1
3	Что означает термин "проверка структурных изменений".	1
4	Смещенными или несмещенными являются ридж-оценки. Поясните ответ.	1
5	Что понимается под помехоустойчивостью внешних критериев	1
6	Покажите, что если P - матрица проектирования на Ω ; где $\Omega = R[X]$ - пространство образов, порожаемое столбцами матрицы X , то $(I - P)^2 = I - P$.	2
7	Как проверить незначимость параметра линейной регрессионной модели	2
8	Что понимается под доступным обобщенным методом наименьших квадратов.	2
9	Покажите, что в среднем норма оценок параметров регрессионной модели при усилении эффекта мультиколлинеарности (например, при стремлении минимального собственного числа к 0) стремится к бесконечности.	2
10	Покажите, что в случае "недобора" регрессоров соответствующие компоненты вектора параметров оцениваются со смещением.	2
11	Какие наблюдения считаются "плохо влияющими".	2
12	Если для одной модели $R^2 = 0.6$, а для второй $R^2 = 0.9$, то во сколько раз остаточная сумма квадратов для второй модели меньше остаточной суммы квадратов для первой модели	2
13	К каким последствиям приводят те или иные ошибки спецификации модели	2
14	Назовите несколько (два – три) внешних критериев, относящихся к группе "критерии согласованности".	2
15	Оценивание параметров в условиях повторных наблюдений	5
16	Задача	12
	Итого	40

Задача

Дана таблица наблюдений:

№	x	y
1	-1	-3.1
2	-1	-2.9
3	0	0.1
4	0	-0.1
5	1	-1.1
6	1	-0.9

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 $\eta_1 = \theta_1 + \theta_2 x^2$ проверить гипотезу $H: \theta_2 = -2$. 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.

Дисциплина "Статистические методы анализа данных"

Билет № 7

№	Вопрос	Баллов
1	Почему этап построения зависимостей под названием "определение класса допустимых решений" называют также этапом параметризации модели.	1
2	Если ошибки наблюдения распределены по нормальному закону, то по какому закону и с какими его параметрами распределены НЛО оценки параметров линейной регрессионной модели и почему.	1
3	Покажите, что дисперсия НЛО оценки равна $D(\hat{\theta}) = \sigma^2 (X^T X)^{-1}$.	1
4	Что означает термин "проверка структурных изменений".	1
5	Как оценить число главных компонент.	1
6	Покажите, что если P - матрица проектирования на Ω ; где $\Omega = R[X]$ - пространство образов, порожаемое столбцами матрицы X , то $(I - P)^2 = I - P$.	2
7	Покажите на простом примере, что матрица $X^T X$ будет бесконечно возрастать при стремлении числа наблюдений к ∞ .	2
8	Что понимается под доступным обобщенным методом наименьших квадратов.	2
9	Покажите, что в среднем норма оценок параметров регрессионной модели при усилении эффекта мультиколлинеарности (например, при стремлении минимального собственного числа к 0) стремится к бесконечности.	2
10	Как узнать какие именно входные факторы ответственны за возникновение эффекта мультиколлинеарности	2
11	Какое наблюдение можно принять за "точку разбалансировки".	2
12	Если для одной модели $R^2 = 0.6$, а для второй $R^2 = 0.9$, то во сколько раз остаточная сумма квадратов для второй модели меньше остаточной суммы квадратов первой модели	2
13	Объясните, почему коэффициент детерминации при добавлении в модель регрессоров ведет себя как возрастающая функция от числа регрессоров.	2
14	Как вводится субидеальный критерий стабильности на основе повторных наблюдений.	2
15	Ридж-оценки неизвестных параметров и их свойства	5
16	Задача	12
	Итого	40

Задача

Дана таблица наблюдений:

№	x	y
1	-1	1.1
2	-1	0.9
3	0	0.1
4	0	-0.1
5	1	3.1
6	1	2.9

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 $\eta_1 = \theta_1 + \theta_2 x^2$ проверить гипотезу $H: \theta_1 = 1$. 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.

Дисциплина "Статистические методы анализа данных"

Билет №8

№	Вопрос	Баллов
1	Почему предпочтительнее использовать симметричные внешние критерии нежели несимметричные	1
2	Для чего в теореме Гаусса-Маркова используется предположение А3: $rg(X) = m$.	1
3	Для чего в теореме Гаусса-Маркова используется предположение А1: $E(\varepsilon) = 0$.	1
4	Какова история происхождения термина "ридж-оценки".	1
5	Как освободить некоторые меры измерения эффекта мультиколлинеарности от влияния масштаба факторов	1
6	Покажите, что $\hat{e}^T X = 0$, где $\hat{e}$ - остатки, X – матрица наблюдений.	2
7	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 для первой модели	2
8	Как можно найти (оценить) неизвестный параметр (коэффициент) автокорреляции ошибок наблюдений.	2
9	Покажите, что в модели, состоящей только из свободного члена, МНК оценка этого параметра равна среднему значению отклика по выборке.	2
10	Покажите, что если P - матрица проектирования на Ω ; где $\Omega = R[X]$ - пространство образов, порожаемое столбцами матрицы X , то $P^2 = P$.	2
11	Получите выражение для оценки неизвестной дисперсии σ^2 по параллельным наблюдениям	2
12	Покажите, что в модели, содержащей свободный член, сумма остатков $e_i = y_i - \hat{y}_i$ равна нулю.	2
13	Назовите недостаток максимальной сопряженности как меры измерения мультиколлинеарности	2
14	Какое наблюдение можно принять за "точку разбалансировки".	2
15	Обобщенный МНК.	5
16	Задача	12
	Итого	40

Задача

Дана таблица наблюдений:

№	x	y
1	-1	4.1
2	-1	3.9
3	0	5.1
4	0	4.9
5	1	6.1
6	1	5.9

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 $H: \theta_2 = 1$. 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.

Дисциплина "Статистические методы анализа данных"

Билет № 9

№	Вопрос	Баллов
1	Дайте толкование выражения "несферические возмущения".	1
2	Обобщенный МНК и ридж-оценивание – это одно и то же?	1
3	В каких целях используются корреляционные поля. Что это за понятие	1
4	Как обнаружить аномальное наблюдение по остатку, используя "правило трех сигм".	1
5	Какое свойство ридж-оценок можно считать главным и почему	1
6	Что понимается под доступным обобщенным методом наименьших квадратов	2
7	Почему абсолютно помехоустойчивый критерий получил такое название.	2
8	Опишите в общих чертах многорядный комбинаторный алгоритм.	2
9	Покажите, что дисперсия НЛО оценки равна $D(\hat{\theta}) = \sigma^2 (X^T X)^{-1}$.	2
10	Покажите, что если P - матрица проектирования на Ω ; где $\Omega = R[X]$ - пространство образов, порожаемое столбцами матрицы, то $(I_n - P)$ - матрица ортогонального проектирования на $\Omega^\perp$.	2
11	Покажите, что в модели, состоящей только из свободного члена, МНК оценка этого параметра равна среднему значению отклика по выборке	2
12	Покажите, что в среднем норма оценок параметров регрессионной модели при усилении эффекта мультиколлинеарности (например, при стремлении минимального собственного числа к 0) стремится к бесконечности.	2
13	Как узнать какие именно входные факторы ответственны за возникновение эффекта мультиколлинеарности.	2
14	Каким недостатком обладает Тест Дарбина-Уотсона.	2
15	Эффективный алгоритм построения всех регрессий.	5
16	Задача	12
	Итого	40

Задача

Дана таблица наблюдений:

№	x	y
1	-1	7.1
2	-1	6.9
3	0	5.1
4	0	4.9
5	1	3.1
6	1	2.9

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 $H: \theta_1 = 0$. 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.

Дисциплина "Статистические методы анализа данных"
Билет № 10

№	Вопрос	Баллов
1	Назовите основные типы прикладных целей исследования зависимостей.	1
2	Почему этап построения зависимостей под названием "определение класса допустимых решений" называют также этапом параметризации модели.	1
3	Что означает термин "линейная оценка".	1
4	Назовите недостатки и достоинства максимальной парной сопряженности как меры измерения мультиколлинеарности	1
5	Перечислите основные меры измерения эффекта мультиколлинеарности	1
6	Как оценить число главных компонент.	2
7	Что понимается под термином "внешние критерии" качества моделей.	2
8	Если наблюдений не очень много, то какой критерий качества моделей целесообразно использовать при выборе модели оптимальной сложности.	2
9	Покажите, что дисперсия НЛО оценки равна $D(\hat{\theta}) = \sigma^2 (X^T X)^{-1}$.	2
10	Получите выражение для оценки неизвестной дисперсии σ^2 по методу максимального правдоподобия	2
11	Проверка данных на автокорреляцию. Тест Дарбина-Уотсона.	2
12	Как запишется гипотеза о проверке незначимости параметра и как ее проверить.	2
13	Какими свойствами должны обладать так называемые главные компоненты.	2
14	Назовите основные свойства оператора выметания. Какова оптимальная последовательность выметаний - так называемый "счетчик Гарсайда".	2
15	Внутренние критерии качества модели. Вывести критерий Мэллоуса	5
16	Задача	12
	Итого	40

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 Мэллоуса? Для вычисления критерия Мэллоуса 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.

Дисциплина "Статистические методы анализа данных"

Билет № 11

№	Вопрос	Баллов
1	Перечислите типовые задачи одномерного и многомерного анализа данных.	1
2	Что означает понятие "линейная по параметрам регрессионная модель".	1
3	Какими свойствами обладает наилучшая линейная оценка.	1
4	Почему предпочтительнее использовать симметричные внешние критерии нежели несимметричные	1
5	Какие меры измерения эффекта мультиколлинеарности свободны от влияния масштаба факторов	1
6	Для чего ведется анализ геометрической структуры исходных данных	2
7	Как запишется гипотеза о проверке незначимости регрессии и как ее проверить	2
8	Что понимается под терминами "Data Mining", "Web Mining". Перечислите основные задачи.	2
9	Покажите на простом примере, что матрица $X^T X$ будет бесконечно возрастать при стремлении числа наблюдений к ∞ .	2
10	Покажите, что $\hat{e}^T X = 0$, где $\hat{e}$ - остатки, X – матрица наблюдений.	2
11	Получите выражение для ковариационной матрицы ошибок наблюдений, если известно, что они описываются процессом авторегрессии первого порядка	2
12	Как узнать какие именно входные факторы ответственны за возникновение эффекта мультиколлинеарности	2
13	Объясните почему коэффициент детерминации при добавлении в модель регрессоров ведет себя как возрастающая функция от числа регрессоров.	2
14	Как вводится субидеальный критерий стабильности на основе повторных наблюдений.	2
15	Наилучшие линейные оценки (НЛО). Теорема Гаусса-Маркова (с доказательством эффективности оценок).	5
16	Задача	12
	Итого	40

Задача

Дана таблица наблюдений:

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

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 $H: \theta_1 = 0$. 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.

Дисциплина "Статистические методы анализа данных"
Билет № 12

№	Вопрос	Баллов
1	Назовите разделы многомерного статистического анализа, которые используются в задачах анализа зависимостей при различных комбинациях типов факторов.	1
2	Какова история происхождения термина "Регрессионный анализ".	1
3	Назовите предположения, которые используются в теореме Гаусса-Маркова.	1
4	Дайте толкование выражения "несферические возмущения".	1
5	Обобщенный МНК и ридж-оценивание – это одно и то же?	1
6	В каких целях используются корреляционные поля. Что это за понятие. Примеры.	2
7	Назовите несколько (два – три) внешних критериев, относящихся к группе "критерии точности".	2
8	Назовите основные этапы статистического исследования зависимостей	2
9	Покажите, что НЛО оценка параметров линейной регрессионной модели является несмещенной	2
10	Покажите, что поскольку $y = \hat{y} + \hat{e}$, то $y^T y = \hat{y}^T \hat{y} + \hat{e}^T \hat{e}$, где $\hat{y} = Py$, $\hat{e} = (I - P)y$.	2
11	Что понимается под доступным обобщенным методом наименьших квадратов	2
12	Как освободить некоторые меры измерения эффекта мультиколлинеарности от влияния масштаба факторов.	2
13	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 для первой модели.	2
14	Дайте интерпретацию гипотезы о структурных изменениях с позиции использование внешних критериев.	2
15	Наилучшие линейные оценки (НЛО). Теорема Гаусса-Маркова (с доказательством состоятельности оценок).	5
16	Задача	12
	Итого	40

Задача

Дана таблица наблюдений:

№	x	y
1	-1	5.1
2	-1	4.9
3	0	3.1
4	0	2.9
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 $\eta_1 = \theta_1 + \theta_2 x^2$ проверить гипотезу $H: \theta_2 = 2$. 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.

Дисциплина "Статистические методы анализа данных"
Билет № 13

№	Вопрос	Баллов
1	Что обозначают термины "пассивный" и "активный" эксперимент.	1
2	Как Вы понимаете термин "адекватность" модели.	1
3	Что означает термин "интервальное оценивание" параметров модели	1
4	Как обнаружить аномальное наблюдение, используя "правило трех сигм".	1
5	Какое свойство ридж-оценок можно считать главным и почему	1
6	Перечислите критерии качества моделей, основанные на использовании одной выборки.	2
7	Почему абсолютно помехоустойчивый критерий получил такое название.	2
8	Опишите в общих чертах многорядный комбинаторный алгоритм	2
9	Покажите на простом примере, что матрица $X^T X$ будет бесконечно возрастать при стремлении числа наблюдений к ∞ .	2
10	Покажите, что если P - матрица проектирования на Ω ; где $\Omega = R[X]$ - пространство образов, порожаемое столбцами матрицы, то $(I_n - P)$ - матрица ортогонального проектирования на $\Omega^\perp$.	2
11	Покажите, что в модели, состоящей только из свободного члена, МНК оценка этого параметра равна среднему значению отклика по выборке	2
12	Покажите, что при усилении эффекта мультиколлинеарности дисперсия оценок параметров линейной регрессионной модели может бесконечно возрастать	2
13	Покажите, что ридж-оценка параметров линейной регрессионной модели является линейным преобразованием МНК оценки	2
14	Поясните, почему следует ожидать, что графики зависимостей внешних критериев качества моделей от числа включенных в модель регрессоров, будут иметь минимум.	2
15	Несмещенное оценивание неизвестной дисперсии ошибки наблюдения	5
16	Задача	12
	Итого	40

Задача

Дана таблица наблюдений:

№	x	y
1	-1	1.1
2	-1	0.9
3	0	3.1
4	0	2.9
5	1	1.1
6	1	0.9

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 $\eta_1 = \theta_1 + \theta_2 x^2$ проверить гипотезу $H: \theta_1 = 0$. 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.

Дисциплина "Статистические методы анализа данных"
Билет № 14

№	Вопрос	Баллов
1	Перечислите известные Вам методы оценивания параметров линейных регрессионных моделей	1
2	Для чего в теореме Гаусса-Маркова используется предположение АЗ: $rg(X) = m$.	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".	2
7	Покажите, что в случае "недобора" регрессоров соответствующие компоненты вектора параметров оцениваются смещено	2
8	Для чего в теореме Гаусса-Маркова используется предположение А1: $E(\varepsilon) = 0$.	2
9	По каким причинам и на каких этапах исследования исследователь вынужден возвращаться на более ранние этапы исследования зависимости.	2
10	Покажите, что если P - матрица проектирования на Ω ; где $\Omega = R[X]$ - пространство образов, порождаемое столбцами матрицы X , то $P^2 = P$.	2
11	Получите выражение для оценки неизвестной дисперсии σ^2 по параллельным наблюдениям	2
12	Покажите, что в модели, содержащей свободный член, сумма остатков $e_i = y_i - \hat{y}_i$ равна нулю.	2
13	Как можно найти (оценить) неизвестный параметр (коэффициент) автокорреляции ошибок наблюдений.	2
14	Какое наблюдение можно принять за "точку разбалансировки".	2
15	Оценивание параметров в модели регрессии при наличии линейных ограничений	5
16	Задача	12
	Итого	40

Задача

Дана таблица наблюдений:

№	x	y
1	-1	-1.1
2	-1	-0.9
3	0	0.1
4	0	-0.1
5	1	3.1
6	1	2.9

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 $\eta_1 = \theta_1 + \theta_2 x^2$ проверить гипотезу $H: \theta_2 = 0$. 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.

Дисциплина "Статистические методы анализа данных"
Билет № 15

№	Вопрос	Баллов
1	Что понимается под термином "структура" модели.	1
2	Если ошибки наблюдения распределены по нормальному закону, то по какому закону распределены НЛО оценки параметров линейной регрессионной модели и почему.	1
3	Что означает термин "проверка структурных изменений".	1
4	Смещенными или несмещенными являются ридж-оценки. Поясните ответ.	1
5	Что понимается под помехоустойчивостью внешних критериев	1
6	Покажите, что если P - матрица проектирования на Ω ; где $\Omega = R[X]$ - пространство образов, порожаемое столбцами матрицы X , то $(I - P)^2 = I - P$.	2
7	Как проверить незначимость параметра линейной регрессионной модели	2
8	Что понимается под доступным обобщенным методом наименьших квадратов.	2
9	Покажите, что в среднем норма оценок параметров регрессионной модели при усилении эффекта мультиколлинеарности (например, при стремлении минимального собственного числа к 0) стремится к бесконечности.	2
10	Покажите, что в случае "недобора" регрессоров соответствующие компоненты вектора параметров оцениваются со смещением.	2
11	Какие наблюдения считаются "плохо влияющими".	2
12	Если для одной модели $R^2 = 0.6$, а для второй $R^2 = 0.9$, то во сколько раз остаточная сумма квадратов для второй модели меньше остаточной суммы квадратов для первой модели	2
13	К каким последствиям приводят те или иные ошибки спецификации модели	2
14	Назовите несколько (два – три) внешних критериев, относящихся к группе "критерии согласованности".	2
15	Оценивание параметров в условиях повторных наблюдений	5
16	Задача	12
	Итого	40

Задача

Дана таблица наблюдений:

№	x	y
1	-1	-3.1
2	-1	-2.9
3	0	0.1
4	0	-0.1
5	1	-1.1
6	1	-0.9

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 $\eta_1 = \theta_1 + \theta_2 x^2$ проверить гипотезу $H: \theta_2 = -2$. 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.

Дисциплина "Статистические методы анализа данных"
Билет № 16

№	Вопрос	Баллов
1	Почему этап построения зависимостей под названием "определение класса допустимых решений" называют также этапом параметризации модели.	1
2	Если ошибки наблюдения распределены по нормальному закону, то по какому закону и с какими его параметрами распределены НЛО оценки параметров линейной регрессионной модели и почему.	1
3	Покажите, что дисперсия НЛО оценки равна $D(\hat{\theta}) = \sigma^2 (X^T X)^{-1}$.	1
4	Что означает термин "проверка структурных изменений".	1
5	Как оценить число главных компонент.	1
6	Покажите, что если P - матрица проектирования на Ω ; где $\Omega = R[X]$ - пространство образов, порожаемое столбцами матрицы X , то $(I - P)^2 = I - P$.	2
7	Покажите на простом примере, что матрица $X^T X$ будет бесконечно возрастать при стремлении числа наблюдений к ∞ .	2
8	Что понимается под доступным обобщенным методом наименьших квадратов.	2
9	Покажите, что в среднем норма оценок параметров регрессионной модели при усилении эффекта мультиколлинеарности (например, при стремлении минимального собственного числа к 0) стремится к бесконечности.	2
10	Как узнать какие именно входные факторы ответственны за возникновение эффекта мультиколлинеарности	2
11	Какое наблюдение можно принять за "точку разбалансировки".	2
12	Если для одной модели $R^2 = 0.6$, а для второй $R^2 = 0.9$, то во сколько раз остаточная сумма квадратов для второй модели меньше остаточной суммы квадратов первой модели	2
13	Объясните, почему коэффициент детерминации при добавлении в модель регрессоров ведет себя как возрастающая функция от числа регрессоров.	2
14	Как вводится субидеальный критерий стабильности на основе повторных наблюдений.	2
15	Ридж-оценки неизвестных параметров и их свойства	5
16	Задача	12
	Итого	40

Задача

Дана таблица наблюдений:

№	x	y
1	-1	1.1
2	-1	0.9
3	0	0.1
4	0	-0.1
5	1	3.1
6	1	2.9

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 $\eta_1 = \theta_1 + \theta_2 x^2$ проверить гипотезу $H: \theta_1 = 1$. 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.

Дисциплина "Статистические методы анализа данных"
Билет №17

№	Вопрос	Баллов
1	Почему предпочтительнее использовать симметричные внешние критерии нежели несимметричные	1
2	Для чего в теореме Гаусса-Маркова используется предположение А3: $rg(X) = m$.	1
3	Для чего в теореме Гаусса-Маркова используется предположение А1: $E(\varepsilon) = 0$.	1
4	Какова история происхождения термина "ридж-оценки".	1
5	Как освободить некоторые меры измерения эффекта мультиколлинеарности от влияния масштаба факторов	1
6	Покажите, что $\hat{\varepsilon}^T X = 0$, где $\hat{\varepsilon}$ - остатки, X – матрица наблюдений.	2
7	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 для первой модели	2
8	Как можно найти (оценить) неизвестный параметр (коэффициент) автокорреляции ошибок наблюдений.	2
9	Покажите, что в модели, состоящей только из свободного члена, МНК оценка этого параметра равна среднему значению отклика по выборке.	2
10	Покажите, что если P - матрица проектирования на Ω ; где $\Omega = R[X]$ - пространство образов, порожаемое столбцами матрицы X , то $P^2 = P$.	2
11	Получите выражение для оценки неизвестной дисперсии σ^2 по параллельным наблюдениям	2
12	Покажите, что в модели, содержащей свободный член, сумма остатков $e_i = y_i - \hat{y}_i$ равна нулю.	2
13	Назовите недостаток максимальной сопряженности как меры измерения мультиколлинеарности	2
14	Какое наблюдение можно принять за "точку разбалансировки".	2
15	Обобщенный МНК.	5
16	Задача	12
	Итого	40

Задача

Дана таблица наблюдений:

№	x	y
1	-1	4.1
2	-1	3.9
3	0	5.1
4	0	4.9
5	1	6.1
6	1	5.9

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 $H: \theta_2 = 1$. 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.

Дисциплина "Статистические методы анализа данных"
Билет № 18

№	Вопрос	Баллов
1	Дайте толкование выражения "несферические возмущения".	1
2	Обобщенный МНК и ридж-оценивание – это одно и то же?	1
3	В каких целях используются корреляционные поля. Что это за понятие	1
4	Как обнаружить аномальное наблюдение по остатку, используя "правило трех сигм".	1
5	Какое свойство ридж-оценок можно считать главным и почему	1
6	Что понимается под доступным обобщенным методом наименьших квадратов	2
7	Почему абсолютно помехоустойчивый критерий получил такое название.	2
8	Опишите в общих чертах многорядный комбинаторный алгоритм.	2
9	Покажите, что дисперсия НЛО оценки равна $D(\hat{\theta}) = \sigma^2 (X^T X)^{-1}$.	2
10	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 $\Omega^\perp$.	2
11	Покажите, что в модели, состоящей только из свободного члена, МНК оценка этого параметра равна среднему значению отклика по выборке	2
12	Покажите, что в среднем норма оценок параметров регрессионной модели при усилении эффекта мультиколлинеарности (например, при стремлении минимального собственного числа к 0) стремится к бесконечности.	2
13	Как узнать какие именно входные факторы ответственны за возникновение эффекта мультиколлинеарности.	2
14	Каким недостатком обладает Тест Дарбина-Уотсона.	2
15	Эффективный алгоритм построения всех регрессий.	5
16	Задача	12
	Итого	40

Задача

Дана таблица наблюдений:

№	x	y
1	-1	7.1
2	-1	6.9
3	0	5.1
4	0	4.9
5	1	3.1
6	1	2.9

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 $H: \theta_1 = 0$. 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.

Дисциплина "Статистические методы анализа данных"
Билет № 19

№	Вопрос	Баллов
1	Назовите основные типы прикладных целей исследования зависимостей.	1
2	Почему этап построения зависимостей под названием "определение класса допустимых решений" называют также этапом параметризации модели.	1
3	Что означает термин "линейная оценка".	1
4	Назовите недостатки и достоинства максимальной парной сопряженности как меры измерения мультиколлинеарности	1
5	Перечислите основные меры измерения эффекта мультиколлинеарности	1
6	Как оценить число главных компонент.	2
7	Что понимается под термином "внешние критерии" качества моделей.	2
8	Если наблюдений не очень много, то какой критерий качества моделей целесообразно использовать при выборе модели оптимальной сложности.	2
9	Покажите, что дисперсия НЛО оценки равна $D(\hat{\theta}) = \sigma^2 (X^T X)^{-1}$.	2
10	Получите выражение для оценки неизвестной дисперсии σ^2 по методу максимального правдоподобия	2
11	Проверка данных на автокорреляцию. Тест Дарбина-Уотсона.	2
12	Как запишется гипотеза о проверке незначимости параметра и как ее проверить.	2
13	Какими свойствами должны обладать так называемые главные компоненты.	2
14	Назовите основные свойства оператора выметания. Какова оптимальная последовательность выметаний - так называемый "счетчик Гарсайда".	2
15	Внутренние критерии качества модели. Вывести критерий Мэллоуса	5
16	Задача	12
	Итого	40

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 Мэллоуса? Для вычисления критерия Мэллоуса 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.

Дисциплина "Статистические методы анализа данных"
Билет № 20

№	Вопрос	Баллов
1	Перечислите типовые задачи одномерного и многомерного анализа данных.	1
2	Что означает понятие "линейная по параметрам регрессионная модель".	1
3	Какими свойствами обладает наилучшая линейная оценка.	1
4	Почему предпочтительнее использовать симметричные внешние критерии нежели несимметричные	1
5	Какие меры измерения эффекта мультиколлинеарности свободны от влияния масштаба факторов	1
6	Для чего ведется анализ геометрической структуры исходных данных	2
7	Как запишется гипотеза о проверке незначимости регрессии и как ее проверить	2
8	Что понимается под терминами "Data Mining", "Web Mining". Перечислите основные задачи.	2
9	Покажите на простом примере, что матрица $X^T X$ будет бесконечно возрастать при стремлении числа наблюдений к ∞ .	2
10	Покажите, что $\hat{e}^T X = 0$, где $\hat{e}$ - остатки, X – матрица наблюдений.	2
11	Получите выражение для ковариационной матрицы ошибок наблюдений, если известно, что они описываются процессом авторегрессии первого порядка	2
12	Как узнать какие именно входные факторы ответственны за возникновение эффекта мультиколлинеарности	2
13	Объясните почему коэффициент детерминации при добавлении в модель регрессоров ведет себя как возрастающая функция от числа регрессоров.	2
14	Как вводится субидеальный критерий стабильности на основе повторных наблюдений.	2
15	Наилучшие линейные оценки (НЛО). Теорема Гаусса-Маркова (с доказательством эффективности оценок).	5
16	Задача	12
	Итого	40

Задача

Дана таблица наблюдений:

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

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 $H: \theta_1 = 0$. 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.

Дисциплина "Статистические методы анализа данных"

Билет №21

№	Вопрос	Баллов
1	Назовите разделы многомерного статистического анализа, которые используются в задачах анализа зависимостей при различных комбинациях типов факторов.	1
2	Какова история происхождения термина "Регрессионный анализ".	1
3	Назовите предположения, которые используются в теореме Гаусса-Маркова.	1
4	Дайте толкование выражения "несферические возмущения".	1
5	Обобщенный МНК и ридж-оценивание – это одно и то же?	1
6	В каких целях используются корреляционные поля. Что это за понятие. Примеры.	2
7	Назовите несколько (два – три) внешних критериев, относящихся к группе "критерии точности".	2
8	Назовите основные этапы статистического исследования зависимостей	2
9	Покажите, что НЛО оценка параметров линейной регрессионной модели является несмещенной	2
10	Покажите, что поскольку $y = \hat{y} + \hat{e}$, то $y^T y = \hat{y}^T \hat{y} + \hat{e}^T \hat{e}$, где $\hat{y} = Py$, $\hat{e} = (I - P)y$.	2
11	Что понимается под доступным обобщенным методом наименьших квадратов	2
12	Как освободить некоторые меры измерения эффекта мультиколлинеарности от влияния масштаба факторов.	2
13	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 для первой модели.	2
14	Дайте интерпретацию гипотезы о структурных изменениях с позиции использование внешних критериев.	2
15	Наилучшие линейные оценки (НЛО). Теорема Гаусса-Маркова (с доказательством состоятельности оценок).	5
16	Задача	12
	Итого	40

Задача

Дана таблица наблюдений:

№	x	y
1	-1	5.1
2	-1	4.9
3	0	3.1
4	0	2.9
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 $\eta_1 = \theta_1 + \theta_2 x^2$ проверить гипотезу $H: \theta_2 = 2$. 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.

Дисциплина "Статистические методы анализа данных"
Билет № 22

№	Вопрос	Баллов
1	Что обозначают термины "пассивный" и "активный" эксперимент.	1
2	Как Вы понимаете термин "адекватность" модели.	1
3	Что означает термин "интервальное оценивание" параметров модели	1
4	Как обнаружить аномальное наблюдение, используя "правило трех сигм".	1
5	Какое свойство ридж-оценок можно считать главным и почему	1
6	Перечислите критерии качества моделей, основанные на использовании одной выборки.	2
7	Почему абсолютно помехоустойчивый критерий получил такое название.	2
8	Опишите в общих чертах многорядный комбинаторный алгоритм	2
9	Покажите на простом примере, что матрица $X^T X$ будет бесконечно возрастать при стремлении числа наблюдений к ∞ .	2
10	Покажите, что если P - матрица проектирования на Ω ; где $\Omega = R[X]$ - пространство образов, порожаемое столбцами матрицы, то $(I_n - P)$ - матрица ортогонального проектирования на $\Omega^\perp$.	2
11	Покажите, что в модели, состоящей только из свободного члена, МНК оценка этого параметра равна среднему значению отклика по выборке	2
12	Покажите, что при усилении эффекта мультиколлинеарности дисперсия оценок параметров линейной регрессионной модели может бесконечно возрастать	2
13	Покажите, что ридж-оценка параметров линейной регрессионной модели является линейным преобразованием МНК оценки	2
14	Поясните, почему следует ожидать, что графики зависимостей внешних критериев качества моделей от числа включенных в модель регрессоров, будут иметь минимум.	2
15	Несмещенное оценивание неизвестной дисперсии ошибки наблюдения	5
16	Задача	12
	Итого	40

Задача

Дана таблица наблюдений:

№	x	y
1	-1	1.1
2	-1	0.9
3	0	3.1
4	0	2.9
5	1	1.1
6	1	0.9

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 $\eta_1 = \theta_1 + \theta_2 x^2$ проверить гипотезу $H: \theta_1 = 0$. 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.

Дисциплина "Статистические методы анализа данных"
Билет № 23

№	Вопрос	Баллов
1	Перечислите известные Вам методы оценивания параметров линейных регрессионных моделей	1
2	Для чего в теореме Гаусса-Маркова используется предположение А3: $rg(X) = m$.	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".	2
7	Покажите, что в случае "недобора" регрессоров соответствующие компоненты вектора параметров оцениваются смещено	2
8	Для чего в теореме Гаусса-Маркова используется предположение А1: $E(\varepsilon) = 0$.	2
9	По каким причинам и на каких этапах исследования исследователь вынужден возвращаться на более ранние этапы исследования зависимости.	2
10	Покажите, что если P - матрица проектирования на Ω ; где $\Omega = R[X]$ - пространство образов, порожаемое столбцами матрицы X , то $P^2 = P$.	2
11	Получите выражение для оценки неизвестной дисперсии σ^2 по параллельным наблюдениям	2
12	Покажите, что в модели, содержащей свободный член, сумма остатков $e_i = y_i - \hat{y}_i$ равна нулю.	2
13	Как можно найти (оценить) неизвестный параметр (коэффициент) автокорреляции ошибок наблюдений.	2
14	Какое наблюдение можно принять за "точку разбалансировки".	2
15	Оценивание параметров в модели регрессии при наличии линейных ограничений	5
16	Задача	12
	Итого	40

Задача

Дана таблица наблюдений:

№	x	y
1	-1	-1.1
2	-1	-0.9
3	0	0.1
4	0	-0.1
5	1	3.1
6	1	2.9

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 $\eta_1 = \theta_1 + \theta_2 x^2$ проверить гипотезу $H: \theta_2 = 0$. 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.

Дисциплина "Статистические методы анализа данных"
Билет № 24

№	Вопрос	Баллов
1	Что понимается под термином "структура" модели.	1
2	Если ошибки наблюдения распределены по нормальному закону, то по какому закону распределены НЛЮ оценки параметров линейной регрессионной модели и почему.	1
3	Что означает термин "проверка структурных изменений".	1
4	Смещенными или несмещенными являются ридж-оценки. Поясните ответ.	1
5	Что понимается под помехоустойчивостью внешних критериев	1
6	Покажите, что если P - матрица проектирования на Ω ; где $\Omega = R[X]$ - пространство образов, порожаемое столбцами матрицы X , то $(I - P)^2 = I - P$.	2
7	Как проверить незначимость параметра линейной регрессионной модели	2
8	Что понимается под доступным обобщенным методом наименьших квадратов.	2
9	Покажите, что в среднем норма оценок параметров регрессионной модели при усилении эффекта мультиколлинеарности (например, при стремлении минимального собственного числа к 0) стремится к бесконечности.	2
10	Покажите, что в случае "недобора" регрессоров соответствующие компоненты вектора параметров оцениваются со смещением.	2
11	Какие наблюдения считаются "плохо влияющими".	2
12	Если для одной модели $R^2 = 0.6$, а для второй $R^2 = 0.9$, то во сколько раз остаточная сумма квадратов для второй модели меньше остаточной суммы квадратов для первой модели	2
13	К каким последствиям приводят те или иные ошибки спецификации модели	2
14	Назовите несколько (два – три) внешних критериев, относящихся к группе "критерии согласованности".	2
15	Оценивание параметров в условиях повторных наблюдений	5
16	Задача	12
	Итого	40

Задача

Дана таблица наблюдений:

№	x	y
1	-1	-3.1
2	-1	-2.9
3	0	0.1
4	0	-0.1
5	1	-1.1
6	1	-0.9

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 $\eta_1 = \theta_1 + \theta_2 x^2$ проверить гипотезу $H: \theta_2 = -2$. 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.

Дисциплина "Статистические методы анализа данных"
Билет № 25

№	Вопрос	Баллов
1	Почему этап построения зависимостей под названием "определение класса допустимых решений" называют также этапом параметризации модели.	1
2	Если ошибки наблюдения распределены по нормальному закону, то по какому закону и с какими его параметрами распределены НЛО оценки параметров линейной регрессионной модели и почему.	1
3	Покажите, что дисперсия НЛО оценки равна $D(\hat{\theta}) = \sigma^2 (X^T X)^{-1}$.	1
4	Что означает термин "проверка структурных изменений".	1
5	Как оценить число главных компонент.	1
6	Покажите, что если P - матрица проектирования на Ω ; где $\Omega = R[X]$ - пространство образов, порожаемое столбцами матрицы X , то $(I - P)^2 = I - P$.	2
7	Покажите на простом примере, что матрица $X^T X$ будет бесконечно возрастать при стремлении числа наблюдений к ∞ .	2
8	Что понимается под доступным обобщенным методом наименьших квадратов.	2
9	Покажите, что в среднем норма оценок параметров регрессионной модели при усилении эффекта мультиколлинеарности (например, при стремлении минимального собственного числа к 0) стремится к бесконечности.	2
10	Как узнать какие именно входные факторы ответственны за возникновение эффекта мультиколлинеарности	2
11	Какое наблюдение можно принять за "точку разбалансировки".	2
12	Если для одной модели $R^2 = 0.6$, а для второй $R^2 = 0.9$, то во сколько раз остаточная сумма квадратов для второй модели меньше остаточной суммы квадратов первой модели	2
13	Объясните, почему коэффициент детерминации при добавлении в модель регрессоров ведет себя как возрастающая функция от числа регрессоров.	2
14	Как вводится субидеальный критерий стабильности на основе повторных наблюдений.	2
15	Ридж-оценки неизвестных параметров и их свойства	5
16	Задача	12
	Итого	40

Задача

Дана таблица наблюдений:

№	x	y
1	-1	1.1
2	-1	0.9
3	0	0.1
4	0	-0.1
5	1	3.1
6	1	2.9

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 $\eta_1 = \theta_1 + \theta_2 x^2$ проверить гипотезу $H: \theta_1 = 1$. 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.

Ответы к билетам
по дисциплине "Статистические методы анализа данных"
для студентов направления 02.03.03 Математическое обеспечение и администрирование
информационных систем, профиль: Математическое и программное обеспечение
информационных технологий

Билет №1

Вопрос 1

Тун 1. Установление самого факта наличия (или отсутствия) статистически значимой связи между откликом и факторами.

Выбор формы связи (т.е. класса допустимых решений F и конкретного вида $\bar{f}(\bar{x}) \in F$) и состава предикторов играет второстепенную роль.

Тун 2. Прогноз (восстановление) неизвестных значений интересующих нас результирующих показателей по заданным значениям факторов x .

Тун 3. Выявление причинных связей между откликами и факторами, частичное управление значениями системы путем регулирования величин входных факторов.

Здесь на первый план выходят задача правильного определения структуры модели.

Вопрос 2

Этап 4 (определение класса допустимых решений).

Главной целью является определение общего вида, структуры искомой связи между $\bar{y}$ и $\bar{x}$, или, другими словами, описание класса функций F , в рамках которого будет производиться поиск конкретного вида интересующей зависимости.

Чаще всего это описание дается в форме некоторого параметрического семейства функций $\bar{f}(\bar{x}, \bar{\theta})$, поэтому этап этот называют также этапом параметризации модели.

Вопрос 3.

Линейные оценки представимы в виде

$$\hat{\theta} = Ty,$$

где $y^T = (y_1, \dots, y_n)$, T – матрица $m \times n$.

Вопрос 4

Достоинство – свободно от влияния масштаба. Недостаток – выявляет только парную коллинеарность.

Вопрос 5

Определитель информационной матрицы, минимальное собственное число, мера обусловленности по Нейману-Голдстейну, максимальная парная сопряженность, максимальная сопряженность.

Вопрос 6

Число главных компонент и определяется по числу относительно больших собственных значений матрицы $X^{*T}X^*$. Недобор суммарной изменчивости может при этом составлять несколько процентов.

Вопрос 7

Критерий называется внешним, если его определение основано на использовании новой информации, «свежих» наблюдениях, не использованных при синтезе модели.

Вопрос 8.

Возможно использование критерия скользящего контроля и критерия регулярности.

Вопрос 9

$$D(\hat{\theta}) = E(\hat{\theta} - E\hat{\theta})(\hat{\theta} - E\hat{\theta})^T = E[(X^T X)^{-1} X^T \varepsilon \varepsilon^T X (X^T X)^{-1}] = \sigma^2 (X^T X)^{-1} \blacksquare$$

Вопрос 10

Для получения ММП оценок для Θ и σ^2 необходимо совместно решить

$$\frac{\partial \log(L)}{\partial \Theta} = -\frac{1}{\sigma^2} X^T (y - X\Theta) = 0$$

$$\frac{\partial \log(L)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (y - X\Theta)^T (y - X\Theta) = 0$$

Решения:

$$\hat{\Theta}_{MML} = (X^T X)^{-1} X^T y, \quad \hat{\sigma}_{MML}^2 = (y - X\hat{\Theta}_{MML})^T (y - X\hat{\Theta}_{MML}) / n.$$

Вопрос 11

Наиболее известный критерий для выявления автокорреляции 1-го порядка – это критерий Дарбина-Уотсона. Для регрессии

$y = X\theta + \varepsilon$ вычисляются остатки $\hat{e} = y - X\hat{\theta}$. По ним вычисляется статистика этого критерия:

$$DW = \sum_{t=2}^n (\hat{e}_t - \hat{e}_{t-1})^2 / \sum_{t=1}^n \hat{e}_t^2.$$

Показано, что

$$DW \cong 2(1 - \hat{\rho}),$$

где $\hat{\rho}$ – коэффициент автокорреляции в регрессии $\hat{e}_t = \rho \hat{e}_{t-1} + v$:

$$\hat{\rho} = \sum_{t=2}^n \hat{e}_t \hat{e}_{t-1} / \sum_{t=2}^n \hat{e}_t^2.$$

Статистика всегда удовлетворяет неравенству

$$0 < DW < 4.$$

При отсутствии автокорреляции статистика DW близка к двум, при положительной автокорреляции она близка к нулю, а при отрицательной к 4.

Вопрос 12

Частный случай 1: проверка равенства отдельного параметра заданной величине c .

Н: $\Theta_j = c$ или Н: $a^T \Theta = c$, где a^T – вектор-строка, в которой на j -ом месте стоит 1, на остальных местах – нули. При $c = 0$ проверяется гипотеза о незначимости параметра.

Обозначим $d_{jj} = ((X^T X)^{-1})_{jj}$ – j -ый диагональный элемент. Тогда

$$F = \frac{(a^T \hat{\Theta} - c)^T (a^T \hat{\Theta} - c)}{\hat{\sigma}^2 d_{jj}} = \frac{(\hat{\Theta}_j - c)^2}{\hat{\sigma}^2 d_{jj}}$$

имеет при гипотезе Н распределение $F_{1,n-m}$. т.е. если $F \leq F_{\alpha,1,n-m}$, где $F_{\alpha,1,n-m}$ - критическая точка, то гипотеза о незначимости не отклоняется.

Вопрос 13

Главные компоненты должны быть ортогональными, суммарная их изменчивость должна быть равна таковой для исходных факторов, сами они определяются как линейные комбинации исходных факторов.

Вопрос 14

Оператор выметания обладает следующими свойствами:

- 1) оператор выметания обратим, так что двукратное применение выметания по одному и тому же (в смысле номера) ведущему элементу матрицы равносильно неприменению этого оператора.
- 2) Выметание коммутативно, так что выметание в порядке k , l -ых элементов равносильно выметанию, проведенному в порядке l , k .

Если $m=4$, то оптимальная последовательность будет (счетчик Гарсайда) следующей

$$S_4 = 1 \ 2 \ 1 \ 3 \ 1 \ 2 \ 1 \ 4 \ 1 \ 2 \ 1 \ 3 \ 1 \ 2 \ 1.$$

Применяя метод индукции, получим

$$S_m = S_{m-1} \ m \ S_{m-1}.$$

Вопрос 15

Мэллоус (Mallows) предложил использовать в качестве оценки для Δ_p статистику

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} + 2p - n,$$

где $\hat{\sigma}^2$ - подходящая оценка для σ^2 , $RSS_p = (y - \hat{y}_p)^T (y - \hat{y}_p)$ - остаточная сумма квадратов.

Часто берут $\hat{\sigma}^2 = \frac{RSS}{(n-m)}$, где m - полное число регрессоров.

Полагая $\hat{\sigma}^2 \approx \sigma^2$, имеем

$$E[C_p] \cong \frac{1}{\sigma^2} E[RSS_p] + 2p - n,$$

далее

$$\begin{aligned} E[RSS_p] &= E[(y - \hat{y}_p)^T (y - \hat{y}_p)] = E[y^T (I_n - P_p) y] = \\ &= \sigma^2 \text{tr}[I_n - P_p] + y_{cp}^T (I_n - P_p) y_{cp} = \sigma^2 (n - p) + SSB_p. \end{aligned}$$

Поэтому

$$E[C_p] \approx (n - p) + \frac{SSB_p}{\sigma^2} + 2p - n = p + \frac{SSB_p}{\sigma^2} = \Delta_p,$$

т.е. C_p является приблизительно несмещённой оценкой для Δ_p .

Задача

-
1. Модель $\eta_1 = \theta_1 + \theta_2 x$
 2. Гипотеза принимается

Билет № 2

Вопрос 1

Одномерный анализ:

Сравнение математических ожиданий;
Сравнение дисперсий;
Оценивание параметров распределений;
Установление закона распределения;
Отбраковка данных.

Многомерный анализ:

Исследование зависимостей между признаками;
Классификация объектов;
Снижение размерности пространства признаков.

Вопрос 2

Функция $\eta(x, \theta)$ линейна по параметрам, если

$$E(y/x) = \eta(x, \theta) = \theta^T f(x),$$

Вопрос 3

Несмещенность, состоятельность, эффективность.

Вопрос 4

Использование симметричных форм критериев всегда целесообразно: во-первых, снижается чувствительность критерия к способу разделения данных на части, и, во-вторых, усредняется влияние помех, действующих на различных частях выборки.

Вопрос 5

Максимальная парная сопряженность, максимальная сопряженность

Вопрос 6

При предварительном анализе характера исследуемых зависимостей ограничиваются в основном изучением “геометрии” парных корреляционных полей. Под корреляционным полем переменных (u, v) понимается графическое представление имеющихся измерений $(u_1, v_1), (u_2, v_2), \dots, (u_n, v_n)$ этих переменных в плоскости (u, v) .

В результате такого анализа обычно получают формулировку нескольких рабочих гипотез об общем виде искомой зависимости.

Вопрос 7

Частный случай 2. Проверка незначимости уравнения регрессии.

Пусть задана линейная модель наблюдения:

$$y_i = \Theta_0 + \Theta_1 x_{1,i} + \dots + \Theta_{m-1} x_{m-1,i} + \varepsilon_i, \quad i = \overline{1, n}$$

и требуется установить является ли регрессия с заданными регрессорами незначимой, т.е. мы хотим проверить гипотезу $H: \Theta_1 = \Theta_2 = \dots = \Theta_{m-1} = 0$ или другими словами сравнить данную модель с моделью среднего $y = \Theta_0$.

Гипотеза Н имеет вид: $A\Theta = 0$, где $A = [0 | I_{m-1}]$ - $(m-1) \times m$ матрица ранга $m-1$.

Применима общая теория с $q = m-1$, $RSS = y^T y - \hat{\Theta}^T X^T y$ и

$$RSS_H = \min_{\Theta_0} \sum_{i=1}^n (y_i - \Theta_0)^2 = \sum_{i=1}^n (y_i - \bar{y})^2 = y^T y - n\bar{y}^2.$$

Вопрос 8

Data Mining (или **Knowledge Discovery In Data** – обнаружение знаний в данных) – исследование и обнаружение машиной в сырых данных скрытых знаний, которые ранее были не известны, нетривиальны, практически полезны, доступны для интерпретации человеком.

Данные: цифровые данные, текст, изображение, видео, звук.

Методы Data Mining находятся на стыке различных направлений информационных технологий: прикладной статистики, баз данных, нейронных сетей, нечетких множеств, генетических алгоритмов, искусственного интеллекта и др.

Основные задачи **Data Mining**: классификация, построение зависимостей, кластеризация, ассоциативные правила, визуализация, поиск исключений.

Вопрос 9

Для линейной модели с двумя факторами при 4 наблюдениях в углах квадрата информационная матрица будет диагональной с диагональю 4. При 8 наблюдениях (повторы) диагональ будет равна 8 и т.д.

Вопрос 10

Условие минимума известно – оно определяется как

$$X^T \underbrace{X \hat{\theta}}_{\hat{y}} = X^T y.$$

Отсюда $X^T (y - X \hat{\theta}) = 0$ и далее $X^T (y - \hat{y}) = 0$, т.е. вектор остатков $\hat{e} = y - \hat{y}$ ортогонален Ω .

Вопрос 11

Поскольку ε_{t-1} и v_t независимы, то

$$E(\varepsilon_t^2) = E(\rho \varepsilon_{t-1} + v_t)^2 = \rho^2 E(\varepsilon_{t-1}^2) + E v_t^2 = \rho^2 E(\varepsilon_{t-1}^2) + \sigma_v^2.$$

Будем предполагать, что ошибки наблюдения ε_t имеют одинаковую дисперсию, т.е. .

$E(\varepsilon_t^2) = E(\varepsilon_{t-1}^2)$. С учетом этого можно получить

$$\sigma_\varepsilon^2 = E(\varepsilon_t^2) = E(\varepsilon_t^2) = E(\varepsilon_0^2) = \sigma_v^2 / (1 - \rho^2).$$

Относительно ковариации можно получить, умножая (3.6) на ε_{t-1} :

$$E(\varepsilon_t \varepsilon_{t-1}) \equiv \text{cov}(\varepsilon_t \varepsilon_{t-1}) = \rho \sigma_\varepsilon^2$$

и далее $\text{cov}(\varepsilon_t \varepsilon_{t-2}) = \rho^2 \sigma_\varepsilon^2, \dots, \text{cov}(\varepsilon_t \varepsilon_{t-m}) = \rho^m \sigma_\varepsilon^2$.

Таким образом, можно окончательно выписать ковариационную матрицу для ε :

$$V = \frac{\sigma_v^2}{1 - \rho^2} \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ . & . & . & \dots & . \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{pmatrix} = \sigma_v^2 \Omega.$$

Вопрос 12

Переменные, которым соответствуют элементы вектора y для минимального собственного числа с значимо отличающимися от 0 значениями и порождают мультиколлинеарность.

Вопрос 13

Монотонность коэффициента детерминации от числа включенных в модель регрессоров легко доказывается, если обратиться к его выражению:

$$R^2 = 1 - \frac{RSS}{\sum (y_i - \bar{y})^2}$$

Вопрос 14

При наличии повторной выборки можно предложить критерий селекции, в основу которого положена идея идеального внешнего критерия. При вычислении идеального внешнего критерия вместо зашумленных значений отклика используются точные (незашумленные) значения. Однако в реальной ситуации эти значения нам не известны. В случае повторной выборки вместо неизвестных значений вектора можно использовать и их оценку в виде выборочного среднего $\bar{y}$:

$$\bar{y}^T = (\bar{y}_A^T, \bar{y}_B^T), \bar{y}_{A_i} = (y_{A_i} + y_{B_i})/2, i = 1, \dots, n_A, \bar{y}_B = \bar{y}_A.$$

С использованием вектора $\bar{y}$ можно предложить следующий субидеальный критерий стабильности:

$$\tilde{S}^2 = \|\bar{y}_A - X_A \hat{\theta}_A\|^2 + \|\bar{y}_B - X_B \hat{\theta}_B\|^2.$$

Вопрос 15

Пусть $\tilde{\theta} = C^T y$ - произвольная несмещенная оценка, т.е.

$$E\tilde{\theta} = E(C^T y) = EC^T (X\theta + \varepsilon) = C^T X\theta = \theta.$$

Таким образом для C должно выполняться

$$C^T X = I. \quad (1.6)$$

Обозначим $H = C^T - (X^T X)^{-1} X^T$. Тогда из (1.6)

$$HX = (C^T - (X^T X)^{-1} X^T)X = C^T X - (X^T X)^{-1} X^T X = C^T X - I = 0,$$

т.е.

$$HX = 0.$$

Определим смещение оценки $\tilde{\theta}$:

$$\tilde{\theta} - \theta = C^T y - \theta = C^T X\theta + C^T \varepsilon - \theta = C^T \varepsilon.$$

Вычислим матрицу ковариаций для оценки $\tilde{\theta}$:

$$\begin{aligned} \text{Cov}(\tilde{\theta}) &\equiv D(\tilde{\theta}) = E(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^T = E(C^T \varepsilon \varepsilon^T C) = \sigma^2 C^T C = \\ &= \sigma^2 [H + (X^T X)^{-1} X^T][H^T + X(X^T X)^{-1}] = \\ &= \sigma^2 [HH^T + HX(X^T X)^{-1} + (X^T X)^{-1} X^T H^T + (X^T X)^{-1} X^T X(X^T X)^{-1}] = \\ &= \sigma^2 HH^T + \sigma^2 (X^T X)^{-1} = \sigma^2 HH^T + D(\hat{\theta}) \end{aligned}$$

Действительно второе слагаемое в полученной сумме есть дисперсия оценки $\hat{\theta}$, поскольку: $\hat{\theta} - \theta = (X^T X)^{-1} X^T \varepsilon$ и

$$D(\hat{\theta}) = E(\hat{\theta} - E\hat{\theta})(\hat{\theta} - E\hat{\theta})^T = E[(X^T X)^{-1} X^T \varepsilon \varepsilon^T X(X^T X)^{-1}] = \sigma^2 (X^T X)^{-1} \blacksquare$$

В силу своего представления матрица HH^T - п.п.о., поэтому $D(\tilde{\theta}) \geq D(\hat{\theta}) \blacksquare$

Для того чтобы $HH^T = 0$ необходимо чтобы $H = 0$. Отсюда получаем, что $C^T = (X^T X)^{-1} X^T$ и $\tilde{\theta} = (X^T X)^{-1} X^T y = \hat{\theta}$ ■

Задача

1. Модель $\eta_1 = \theta_1 + \theta_2 x$
2. Гипотеза отвергается

Билет №3

Вопрос 1

Методы статистического исследования зависимостей составляют содержание отдельных частей многомерного статистического анализа.

Они формировались и развивались с учетом специфики анализируемых моделей, обусловленной природой изучаемых переменных.

Соответствующая спецификация этих разделов отражена в следующей таблице.

№	Природа откликов	Природа факторов	Название раздела многомерного статистического анализа
1.	Количественная	Количественная	Регрессионный анализ
2.	Количественная	Одна переменная – время	Анализ временных рядов
3.	Количественная	Качественная	Дисперсионный анализ
4.	Количественная	Смешанная	Ковариационный анализ
5.	Неколичественная	Неколичественная	Анализ ранговых корреляций
6.	Неколичественная или отсутствие	Количественная	Дисперсионный анализ, кластерный анализ
7.	Смешанная	Смешанная	Аппарат логических решающих функций

Вопрос 2

Термины «регрессионная модель», «регрессионный анализ» – устоявшиеся, хотя и не совсем удачные.

Термин «регрессия» ввел антрополог Гальтон (1822-1911) в 1885г. Он установил «закон регрессии». Имеется в виду, что «возврат» (регрессия) потомков в отношении их индивидуальных признаков к родителям происходит в «среднем». В частности он рассматривал размер семян и рост детей и родителей. У рослых родителей не обязательно еще более рослые дети и наоборот, у малорослых родителей – еще более малорослые. Он сравнивал средний рост детей $\bar{Y}$ со средним ростом их родителей $\bar{X}$. Гальтон заметил, что рост детей у высоких (низких) родителей обычно также выше (ниже) среднего роста популяции $\mu \cong \bar{X} \cong \bar{Y}$, но при этом отклонение от μ у детей меньше, чем у родителей. Другими словами, экстремумы в следующем поколении сглаживаются, происходит возвращение назад (*регрессия*) к середине. Гальтон показал, что зависимость $\bar{Y}$ от $\bar{X}$ хорошо выражается уравнением $\bar{Y} - \bar{Y} = (2/3)(\bar{X} - \bar{X})$.

Вопрос 3

A1: $E(\varepsilon) = 0$; A2: $E(\varepsilon\varepsilon^T) = \sigma^2 I_n$; A3: $rg(X) = m$.

Вопрос 4

Термин "сферичность" возмущений происходит из того наблюдения, что в многомерном пространстве распределения помехи ε при $Var(\varepsilon) = \sigma^2 I_n$ (случай равнооточных наблюдений) точки с одинаковой вероятностью образуют сферы. В рассматриваемом случае неравнооточных наблюдений вместо сфер будут образовываться многомерные эллипсоиды.

Вопрос 5

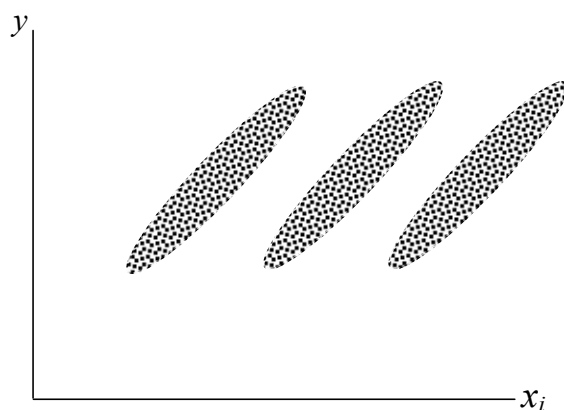
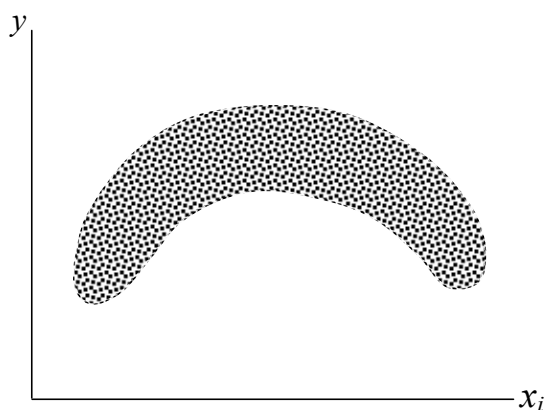
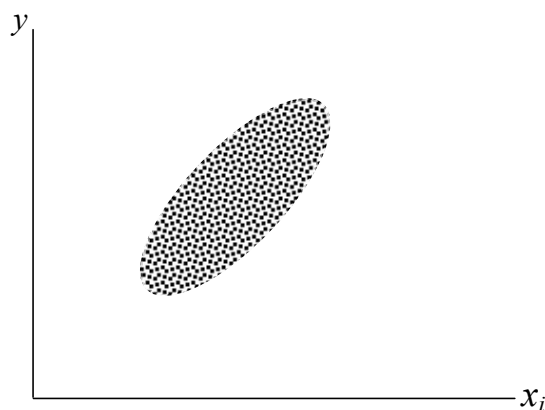
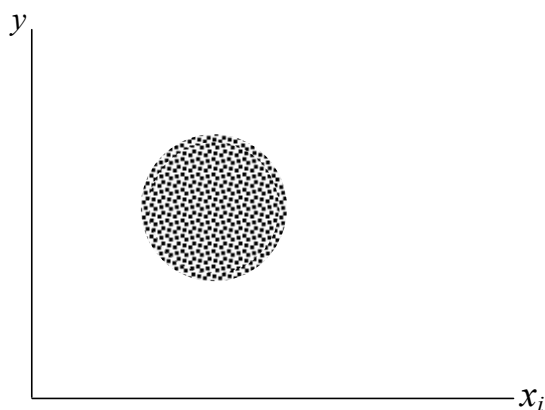
Нет.

Вопрос 6

При предварительном анализе характера исследуемых зависимостей ограничиваются в основном изучением "геометрии" парных корреляционных полей. Под корреляционным полем переменных (u, v) понимается графическое представление имеющихся измерений $(u_1, v_1), (u_2, v_2), \dots, (u_n, v_n)$ этих переменных в плоскости (u, v) .

Анализ парных корреляционных полей состоит обычно в следующем:

- построение на основании имеющихся исходных данных корреляционных полей для всевозможных пар переменных вида (x_j, x_i) и (x_j, y) — всего таких пар будет $k(k+1)/2$. Этот процесс легко автоматизируется с помощью средств современных ЭВМ;
- визуальное прослеживание характера вытянутости каждого корреляционного поля (эллипсоидально-линейное, нелинейное-монотонное, с наличием экстремумов и т. п.);



Это примеры корреляционных полей для пассивного эксперимента.

Вопрос 7

Критерий стабильности, регулярности.

Вопрос 8

Этапы: постановочный. Информационный, корреляционный анализ, определение класса допустимых решений, анализ мультиколлинеарности, оценка параметров, проверка адекватности

Вопрос 9

Имеем

$$\hat{\theta} = (X^T X)^{-1} X^T (X\theta + \varepsilon) = \theta + (X^T X)^{-1} X^T \varepsilon;$$

$$E(\hat{\theta}) = E(\theta) + (X^T X)^{-1} X^T E(\varepsilon) = \theta$$

поскольку действует предположение А1.

Вопрос 10

Поскольку $\hat{y}$ и $\hat{e}$ из ортогональных подпространств, то ясно, что $\hat{y}^T \hat{e} = 0$. Далее, поскольку $y = \hat{y} + \hat{e}$, то $y^T y = \hat{y}^T \hat{y} + \hat{e}^T \hat{e}$.

Вопрос 11

Доступный обобщенный МНК. В случае обнаружения эффекта гетероскедастичности можно провести коррекцию оценок параметров модели с помощью взвешенного МНК с весами $\hat{\sigma}_i^{-2} = h^{-1}(z_i^T \hat{\alpha})$. Может оказаться, что для некоторых t $h^{-1}(z_i^T \hat{\alpha}) < 0$. Если число таких наблюдений невелико, то их можно просто выбросить, в противном случае использовать представление $\sigma_i^2 = \exp(h(z_i^T \alpha))$.

Вопрос 12

Можно рекомендовать использовать информационную матрицу, нормированную на след:
 $X^T X / \text{tr} X^T X$.

Вопрос 13

В 2 раза.

Вопрос 14

Предположим, что ошибки наблюдения распределены по нормальному закону $N(0, \sigma^2 I_N)$. Пусть априорные данные о параметрах модели определены информацией из выборки A и состоят в выполнении $X_A^T X_A \theta = X_A^T y_A$. Поступают новые данные в виде выборки B . По ней производится оценивание параметров и осуществляется проверка гипотезы $H: X_A^T X_A \theta = X_A^T y_A$.

Обозначим

$$RSS = \|y_B - \hat{y}_B(B)\|^2 = (y_B - X_B \hat{\theta}_B)^T (y_B - X_B \hat{\theta}_B) \equiv \varepsilon^2(B), \quad)$$

$$RSS_H = (y_B - X_B \hat{\theta}_H)^T (y_B - X_B \hat{\theta}_H), \quad)$$

где $\hat{\theta}_H$ - оценки, полученные при ограничениях $X_A^T X_A \theta = X_A^T y_A$.

В нашем случае при условии полного ранга матрицы $X_A^T X_A$ имеем

$$\hat{\theta}_H = \hat{\theta}_A = (X_A^T X_A)^{-1} X_A^T y_A.$$

Статистика F для проверки гипотезы H будет следующая

$$F = \frac{(RSS_H - RSS) / p}{RSS / (N_B - p)},$$

где p – число параметров в модели. С учетом ранее введенных обозначений F -статистику можно переписать

$$F = \frac{(\Delta_{reg}^2(B) - \varepsilon^2(B)) / p}{\varepsilon^2(B) / (N_B - p)} = \frac{\Delta_{reg}^2(B)}{\varepsilon^2(B)} - \frac{N_B - p}{p}.$$

Вопрос 15

Под состоятельностью понимается сходимость по вероятности

$$\lim_{N \rightarrow \infty} P \left[\left(\hat{\theta}_N - \theta \right)^T \left(\hat{\theta}_N - \theta \right) \geq \gamma \right] = 0.$$

Состоятельность оценки (точнее – последовательности оценок) означает концентрацию вероятностной массы около истинного значения параметра с ростом объема выборки. В многомерном случае требуется, чтобы состоятельность выполнялась для каждого компонента θ_i вектора θ .

Будем рассматривать состоятельность в среднем квадратичном:

$$E(\hat{\theta}_N - \theta)^T (\hat{\theta}_N - \theta) \rightarrow 0, N \rightarrow \infty.$$

Сходимость в среднеквадратичном влечет сходимость по вероятности. Это сразу вытекает из неравенства Чебышева

$$P \left[\left| \hat{\theta}_N - \theta \right| \geq \gamma \right] \leq \frac{1}{\gamma} E(\hat{\theta}_N - \theta)^2.$$

Величина $E(\hat{\theta}_N - \theta)^T (\hat{\theta}_N - \theta)$ есть след матрицы ковариаций $E(\hat{\theta}_N - \theta)(\hat{\theta}_N - \theta)^T \equiv D(\hat{\theta}_N)$.

Таким образом для выполнения условия

$tr D(\hat{\theta}_N) \equiv E(\hat{\theta}_N - \theta)^T (\hat{\theta}_N - \theta) \rightarrow 0, N \rightarrow \infty$ необходимо потребовать выполнение условия $D(\hat{\theta}_N) \rightarrow O, N \rightarrow \infty$.

Поскольку $D(\hat{\theta}_N) = \sigma^2 \left(\sum_{i=1}^N f(x_i) f^T(x_i) \right)^{-1} = \sigma^2 (X^T X)^{-1}$ необходимо чтобы матрица $X^T X$

бесконечно возрастала с ростом N . Поскольку $f(x_i) f^T(x_i)$ – п.п.о., то всегда найдется такое

k , что $\sigma^2 \sum_{i=1}^N f(x_i) f^T(x_i) < \sigma^2 \sum_{i=1}^{N+k} f(x_i) f^T(x_i)$,

что означает неограниченный рост $X^T X$ ■

Задача

1. Модель $\eta_1 = \theta_1 + \theta_2 x^2$
2. Гипотеза принимается

Билет №4

Вопрос 1

При активном эксперименте исследователь сам устанавливает необходимые комбинации значений входных факторов.

Вопрос 2

Адекватность модели – способность описывать поведение объекта (системы) в среднем в условиях того, что экспериментальные наблюдения за откликом подвержены влиянию случайных помех.

Вопрос 3

Доверительное оценивание для отдельного параметра

Доверительный интервал для отдельного параметра также можно получить исходя из общего выражения:

$$(\Theta_j - \hat{\Theta}_j)^T \frac{1}{d_{jj}} (\Theta_j - \hat{\Theta}_j) \leq \hat{\sigma}^2 F_{\alpha, 1, n-m}$$

Учитывая, что $((X^T X)^{-1})_{jj} \sigma^2 = \sigma^2 (\hat{\Theta}_j)$ - дисперсия оценки отдельного параметра, и что ,

где $t_{\alpha/2, f_R}$ - t -критерий Стьюдента, $f_R = n-m$, то можно записать в виде: $\frac{|\Theta_j - \hat{\Theta}_j|}{\sigma(\hat{\Theta}_j)} \leq t_{\alpha/2, f_R}$.

Или в виде двухстороннего неравенства:

$$\hat{\Theta}_j - t_{\alpha/2, f_R} \sigma(\hat{\Theta}_j) \leq \Theta_j \leq \hat{\Theta}_j + t_{\alpha/2, f_R} \sigma(\hat{\Theta}_j).$$

Вопрос 4

Выделяющиеся точки графика: некоторые из остатков могут по абсолютной величине сильно превосходить все остальные остатки. Если максимальное значение остатка $|d_{max}| > \delta$, где δ - некоторое заранее выбранное число, то наблюдение, имеющее такой остаток, является аномальным. Величина остатка аномального наблюдения зависит от объема выборки n : при $n > 20$ его можно обнаружить по правилу «трех сигм» ($\delta = 3$), при $n < 20$, например, по критерию Смирнова - Граббса. В этом случае вычисляемая статистика имеет вид

$$D = (d_{\max} - \bar{d}) / S_d, \text{ где } \bar{d} - \text{среднее значение остатков, } S_d^2 = 1/n \left(\sum_{i=1}^n (d_i - \bar{d})^2 \right).$$

Вопрос 5

Ридж-оценка имеет минимальную длину в классе оценок с заданным значением суммы квадратов отклонений (RSS).

Вопрос 6

Статистика Мэллоуса, коэффициент детерминации, критерий Акаике.

Вопрос 7

Для критерия Δ_{abs}^2 , например, его среднее значение будет определяться величиной

$$\bar{\Delta}_{abs}^2 = u^T D_{\Delta_{abs}(p)} u + \sigma^2 \cdot p.$$

С увеличением σ^2 минимум критерия в его зависимости от сложности модели смещается влево, т.е. выбираются все более простые модели. Абсолютно помехоустойчивый критерий получил свое название по причине того, что второе слагаемое для него менее всего возрастает при увеличении дисперсии помехи. Другими словами, при увеличении дисперсии помехи минимум данного критерия с наименьшей скоростью смещается влево в сторону все более простых моделей.

Вопрос 8

На первом шаге выбирается некоторое число лучших моделей, состоящих из одного регрессора. На втором этапе к этим моделям добавляются различные регрессоры по одному, и отбираются модели с двумя аргументами, улучшающие величину критерия, и т.д. до его ухудшения или до построения полной модели.

На последнем шаге можно из оставшихся лучших моделей выбрать одну по другому критерию.

Вопрос 9

Для линейной модели с двумя факторами при 4 наблюдениях в углах квадрата информационная матрица будет диагональной с диагональю 4. При 8 наблюдениях (повторы) диагональ будет равна 8 и т.д.

Вопрос 10

Имеем $X^T(y - X\hat{\theta}) = 0$ и далее $X^T(y - \hat{y}) = 0$, т.е. вектор остатков $\hat{e} = y - \hat{y}$ ортогонален Ω . В тоже время справедливо представление $y = \hat{y} + \hat{e}$. Становится понятным, что $\hat{y}$ есть проекция y на $R(X)$, а именно $\hat{y} = Py$. Можно рассматривать также подпространство $R^\perp(X)$ – ортогональное дополнение для $R(X)$ в $R(X)$. Поскольку

$$\hat{e} = y - \hat{y} = y - X\hat{\theta} = (I_n - X(X^T X)^{-1} X^T)y = (I_n - P)y,$$

то $M = I_n - P$ – есть матрица проектирования вектора y на подпространство $R^\perp(X)$.

Вопрос 11

$$\hat{\Theta} = (X^T X)^{-1} X^T y. \quad \hat{\Theta} = (n)^{-1} X^T y = \sum_{i=1}^n y_i / n$$

Вопрос 12

Для наглядности рассмотрим случай $k=2$.

$$y = \theta_1 x_1 + \theta_2 x_2 + \varepsilon.$$

Пусть для удобства мы пронормировали вектора

$$\underline{x}_1^T = (x_{11}, \dots, x_{1n}), \quad \underline{x}_2^T = (x_{21}, \dots, x_{2n}): \quad \underline{x}_1' = \frac{\underline{x}_1}{\|\underline{x}_1\|}, \quad \underline{x}_2' = \frac{\underline{x}_2}{\|\underline{x}_2\|}.$$

Тогда

$$X^T X = \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}, \text{ где } r_{12} = \cos(\underline{x}_1, \underline{x}_2) = \sum \frac{x_{1t} x_{2t}}{\sqrt{\sum x_{1t}^2 \cdot \sum x_{2t}^2}}.$$

Тогда

$$(X^T X)^{-1} = \begin{pmatrix} \frac{1}{(1-r_{12}^2)} & -\frac{r_{12}}{(1-r_{12}^2)} \\ -\frac{r_{12}}{(1-r_{12}^2)} & \frac{1}{(1-r_{12}^2)} \end{pmatrix} \text{ и } \sigma^2(\hat{\theta}_1) = \sigma^2(\hat{\theta}_2) = \frac{\sigma^2}{(1-r_{12}^2)}.$$

Если между переменными x_1 и x_2 существует некая линейная зависимость, т.е. r_{12}^2 близко к 1, то, как следует из последнего выражения, дисперсии оценок параметров будут иметь большие значения. При $r_{12} \rightarrow 1 \Rightarrow \sigma^2(\hat{\theta}_1) \rightarrow \infty$.

Вопрос 13

- 2) $\hat{\beta}(\lambda)$ - является линейным преобразованием МНК-оценки и является смещённой;
Действительно,

$$Z^T y = (I + \lambda(Z^T Z)^{-1})^{-1} (Z^T Z)^{-1} Z^T y = B\hat{\beta},$$

где $B = (I + \lambda(Z^T Z)^{-1})^{-1}$ - матрица линейного преобразования и $B \neq I$, что приводит к смещению оценок $\hat{\beta}(\lambda)$ в отличие от $\hat{\beta}$, которые несмещены.

Вопрос 14

При переусложнении модели имеется дополнительная возможность описывать не только среднее поведение системы, но и флуктуации, обусловленные влиянием помех. На новых тестовых данных помеха уже будет другая, что и создаст условия увеличения ошибки предсказания.

Вопрос 15

Введем в рассмотрение вектор остатков: -

$$\hat{e} = y - \hat{y} = y - X\hat{\theta} = (I_n - X(X^T X)^{-1} X^T) y = (I_n - P)y,$$

где P - матрица проектирования: $\hat{y} = Py$, т.е. y проектируется на Ω ; $(I_n - P)$ - матрица ортогонального проектирования на $\Omega^\perp$, где $\Omega = R[X]$ - пространство образов, порожаемое столбцами матрицы X .

Матрицы P и $I_n - P$ симметричны и идемпотентны, т.е. $P^2 = P$, $(I_n - P)^2 = (I_n - P)$.

Поскольку матрица $I_n - P$ симметрична и идемпотентна, то $\text{rang}[I_n - P] = \text{tr}[I_n - P] = n - \text{tr}[P]$ т.е. у идемпотентной матрицы ранга r имеется r собственных чисел равных 1, а остальные $n-r$ равны 0. В нашем случае:

$$\text{tr}[P] = \text{tr}[X(X^T X)^{-1} X^T] = \text{tr}[X^T X (X^T X)^{-1}] = \text{tr}[I_m] = m.$$

Введем в рассмотрение статистику

$$\hat{e}^T \hat{e} = (y - X\hat{\theta})^T (y - X\hat{\theta}) = y^T (I_n - P)^T (I_n - P) y = y^T (I_n - P)^2 y.$$

Теорема о математическом ожидании квадратичной формы от случайной величины:

Пусть Z - случайный вектор размера $n \times 1$, A - симметричная $n \times n$ матрица. Если $E(Z) = \xi$ и $D(Z) = \Sigma$, то $E[Z^T A Z] = \text{tr}[A \Sigma] + \xi^T A \xi$.

Применительно к нашему случаю имеем:

$$E[\hat{e}^T \hat{e}] = E[y^T (I_n - P) y] = \sigma^2 \text{tr}(I_n - P) + \theta_{\epsilon \text{н}0}^T X^T (I_n - P) X \theta_{\epsilon \text{н}0} = \sigma^2 (n - m).$$

Примечание. Второе слагаемое здесь равно нулю поскольку $(I_n - P)X = 0$.

Отсюда заключаем, что статистика $\hat{\sigma}^2 = \frac{\hat{e}^T \hat{e}}{n - m}$ - несмещенная оценка для σ^2 .

Задача

1. Модель $\eta_1 = \theta_1 + \theta_2 x^2$

2. Гипотеза отвергается

Билет №5

Вопрос 1

НЛО, МНК, ММП, Ридж-оценивание, ОМНК

Вопрос 2

Для возможности обращения информационной матрицы

Вопрос 3

Для ее вычисления необходимо обратить матрицу коэффициентов корреляций, которая может быть плохо обусловлена в условиях мультиколлинеарности.

Вопрос 4

Ридж-оценки впервые были введены А. Гоэрлом при исследовании регрессии 2-го порядка (ridge – гребень, хребет). Название "ридж-оценки" навеяно следующими наблюдениями. При сильной мультиколлинеарности линии равного уровня доверительного эллипсоида для разного уровня значимости (доверительной вероятности) становятся сильно вытянутыми и визуально похожими на хребет на географической карте. Данные оценки использовались именно в условиях мультиколлинеарности и поэтому получили название ридж-оценок.

Вопрос 5

Ошибка спецификации модели – это недобор или перебор регрессоров.

Вопрос 6

Критерии точности.

Типичным критерием этого вида является критерий регулярности:

$$\begin{aligned}\Delta_{reg}^2(B) &= \Delta^2(B/A) = \|y_B - \hat{y}_B(A)\|^2 = \\ &= (y_B - X_B \hat{\theta}_A)^T (y_B - X_B \hat{\theta}_A) = \|y_B - X_B \hat{\theta}_A\|^2\end{aligned}$$

где запись $\Delta^2(B/A)$ означает «ошибка на выборке B модели, коэффициенты которой получены на A », $\hat{\theta}_A = (X_A^T X_A)^{-1} X_A^T y_A$, $\hat{y}_B(A) = X_B \hat{\theta}_A$. Этот критерий несимметричен.

Вопрос 7

Случай недобора.

Пусть истинное уравнение (4.2), а мы оцениваем регрессию

$$y_u = \gamma_1 \omega_{u1} + \dots + \gamma_k \omega_{uk} + \xi_u \quad (4.5)$$

или

$$y = W\gamma + \xi, \quad X = \begin{bmatrix} W & V \end{bmatrix}$$

Другими словами, $\omega_1 = x_1, \dots, \omega_k = x_k$ входят в регрессию (4.5), а остальные $m - k$ переменных $\omega_{k+1}, \dots, \omega_{m-k} = x_m$ в регрессии отсутствуют.

Докажем, что тогда используемая оценка

$$\hat{\gamma} = (W^T W)^{-1} W^T y \quad (4.6)$$

в общем случае является смещённой в том смысле, что $E\hat{\gamma} \neq \alpha_1$.

Действительно,

$$E\hat{\gamma} = (W^T W)^{-1} W^T E y = (W^T W)^{-1} W^T X \alpha = (W^T W)^{-1} W^T [W \ V] \cdot \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix},$$

где α_1 и α_2 - подвекторы вектора α размерности k и $m - k$ соответственно.

Перемножая, получим

$$E\hat{\gamma} = \begin{bmatrix} I_k & (W^T W)^{-1} W^T V \end{bmatrix} \cdot \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \alpha_1 + (W^T W)^{-1} W^T V \alpha_2 \quad \blacksquare$$

Оценка $\hat{\gamma}$ будет несмещённой, когда матрицы W и V ортогональны, то есть $W^T V = 0$.

Можно показать также, что МНК оценка в случае недобора является несостоятельной.

Вопрос 8

Для доказательства несмещенности оценок.

Вопрос 9

По причинам неадекватности модели, обнаружения мультиколлинеарности.

Вопрос 10

$$P = X(X^T X)^{-1} X^T. \quad P^2 = X(X^T X)^{-1} X^T X(X^T X)^{-1} X^T = P$$

Вопрос 11

Точечную оценку дисперсии ошибки эксперимента в случае, когда имеется повторные опыты лучше всего определять через сумму (2.23) по формуле

$\hat{\sigma}_E^2 = SS_E / f_E$, где f_E число степеней свободы суммы SS_E , равное

$f_E = N - N_c = \sum_{v=1}^{N_c} (r_v - 1)$. Оценка $\hat{\sigma}_E^2$ есть не что иное, как объединенная оценка

выборочных дисперсий ошибки опыта, которые можно найти по результатам отдельных серий повторных опытов. Оценка дисперсии ошибки в v -ой серии повторных опытов равна

$$\hat{\sigma}_{Ev}^2 = \frac{SS_{Ev}}{f_{Ev}}, \text{ где } SS_{Ev} = \sum_{j=1}^{r_v} (y_{vj} - \bar{y}_v)^2, \quad f_{Ev} = r_v - 1.$$

Отсюда следует, что объединенная оценка по всем сериям имеет вид

$$\hat{\sigma}_E^2 = \sum_{v=1}^{N_c} \frac{f_{Ev}}{f_E} \hat{\sigma}_{Ev}^2 = \sum_{v=1}^{N_c} SS_{Ev} / f_E = SS_E / f_E.$$

Вопрос 12

. Рассмотрим случай когда в модели присутствует свободный член. В этом случае сумма остатков равно нулю, т.е.

$$\sum (y_i - \hat{y}_i) = 0. \quad (a)$$

Покажем это. Для этого сумму квадратов отклонений для модели со свободным членом

$$\sum (y_i - \theta_0 - \theta_1 x_{1i} - \dots - \theta_{m-1} x_{m-1,i})^2$$

продифференцируем по θ_0 :

$$\sum (y_i - \hat{\theta}_0 - \hat{\theta}_1 x_{1i} - \dots - \hat{\theta}_{m-1} x_{m-1,i}) = 0 \text{ или } \sum (y_i - \hat{y}_i) = 0.$$

Вопрос 13

Ясно, что в реальных ситуациях редко бывает так, что параметр автокорреляции ρ известен. При неизвестном ρ процедуры оценивания параметров θ носят итеративный характер.

Процедура Кохрейна-Оркатта.

1. Для исходной модели оцениваются $\hat{\theta}$ и $\hat{e}$;
2. Рассматривается регрессия $\hat{e}_i = \rho \hat{e}_{i-1} + v_i$ и вычисляется оценка $\hat{\rho}$;
3. Проводится ОМНК оценивание исходной модели с $\hat{\Omega}^{-1}$, где в качестве ρ используется его оценка $\hat{\rho}$;
4. Вычисляется новый вектор остатков $\hat{e}$;
5. Процедура повторяется с п.2.

Останов производится после нескольких итераций или после того как очередное $\hat{\rho}$ мало отличается от предыдущего.

Процедура Хилдрета-Лу.

По схеме ОМНК производится решение следующей экстремальной задачи:

$$\hat{\rho} = \underset{\rho \in [-1, +1]}{\text{Arg min}} \text{RSS}.$$

Попутно при этом мы получаем оценки для вектора параметров θ .

Вопрос 14

$$p_{ii} = f^T(x_i)(X^T X)^{-1} f(x_i)$$

. На практике руководствуются следующими правилами. Значения $p_{ii} \leq 0,2$ позволяют считать надежными данные наблюдения, значения $0,2 \leq p_{ii} \leq 0,5$ относят данное наблюдение к классу рискованных, а при значениях $p_{ii} > 0,5$ i -е наблюдение можно принять за точку разбалансировки.

Вопрос 15

Пусть как и ранее модель наблюдения имеет вид:

$$y = X\Theta + \varepsilon,$$

где X - матрица $n \times m$ ранга m .

Предположим, что мы хотим найти оценки параметра Θ , минимизируя $\varepsilon^T \varepsilon$, при наличии линейных ограничений $A\Theta = c$, где A - известная $q \times m$ матрица ранга $q < m$, а c - известный вектор $q \times 1$.

Применим метод множителей Лагранжа. Рассмотрим целевую функцию:

$$r = \varepsilon^T \varepsilon + (\Theta^T A^T - c^T) \lambda,$$

где $\lambda^T = (\lambda_1, \dots, \lambda_q)$ - вектор множителей Лагранжа. Решим уравнения:

$$A\Theta = c \tag{2.1}$$

$$\frac{\partial r}{\partial \Theta} = -2X^T y + 2X^T X \Theta + A^T \lambda = 0. \quad (2.2)$$

Решения уравнений (2.1) и (2.2) обозначим через $\hat{\lambda}_H$ и $\hat{\Theta}_H$. Тогда из (2.2) получаем:

$$\hat{\Theta}_H = (X^T X)^{-1} X^T y - \frac{1}{2} (X^T X)^{-1} A^T \hat{\lambda}_H = \hat{\Theta} - \frac{1}{2} (X^T X)^{-1} A^T \lambda_H, \quad (2.3)$$

а из (2.1):

$$c = A \hat{\Theta}_H = A \hat{\Theta} - \frac{1}{2} A (X^T X)^{-1} A^T \hat{\lambda}_H.$$

Поскольку матрица $(X^T X)^{-1}$ - п.п.о., то матрица $A(X^T X)^{-1} A^T$ будет п.п.о. ранга q . Следовательно

$$-\frac{1}{2} \hat{\lambda}_H = [A(X^T X)^{-1} A^T]^{-1} (c - A \hat{\Theta}).$$

Окончательно подставляя в (2.3):

$$\hat{\Theta}_H = \hat{\Theta} + (X^T X)^{-1} A^T [A(X^T X)^{-1} A^T]^{-1} (c - A \hat{\Theta}). \quad (2.4)$$

Задача

1. Модель $\eta_1 = \theta_1 x + \theta_2 x^2$
2. Гипотеза отвергается

Билет №6

Вопрос 1

Структура модели – это набор регрессоров.

Вопрос 2

Оценки параметров распределены по нормальному закону с мат.ожиданием равным истинным значениям параметров и дисперсией $\sigma^2 (X^T X)^{-1}$

Вопрос 3

Это проверка того, что модель системы не изменилась со временем.

Вопрос 4

- 2) $\hat{\beta} \lambda$ - является линейным преобразованием МНК-оценки и является смещённой; Действительно,

$$Z^T y = (I + \lambda (Z^T Z)^{-1})^{-1} (Z^T Z)^{-1} Z^T y = B \hat{\beta},$$

где $B = (I + \lambda(Z^T Z)^{-1})^{-1}$ - матрица линейного преобразования и $B \neq I$, что приводит к смещению оценок $\hat{\beta}(\lambda)$ в отличие от $\hat{\beta}$, которые несмещены.

Вопрос 5

Под помехоустойчивостью внешнего критерия будем понимать его способность выбирать из всего множества моделей, построенных по зашумленным данным, такую модель оптимальной сложности, которая достаточно точно восстанавливает не зашумленный выход объекта.

Вопрос 6

$$I - P = I - X(X^T X)^{-1} X^T.$$

$$(I - P)^2 = I - 2P + P^2 =$$

$$= I - 2X(X^T X)^{-1} X^T + X(X^T X)^{-1} X^T X(X^T X(X^T X)^{-1} X^T)^{-1} X^T = I - P$$

Вопрос 7

Частный случай 1: проверка незначимости отдельного параметра $H: \theta_j = c$ или $H: a^T \theta = c$, где a^T - вектор-строка, в которой на j -ом месте стоит 1, на остальных местах - нули. Обозначим $d_{jj} = ((X^T X)^{-1})_{jj}$ - j -ый диагональный элемент. Тогда

$$F = \frac{(a^T \hat{\theta} - c)^2}{\hat{\sigma}^2 d_{jj}} = \frac{(\hat{\theta}_j - c)^2}{\hat{\sigma}^2 d_{jj}}$$

имеет при гипотезе H распределение $F_{1, n-m}$. т.е. если $F \leq F_{\alpha, 1, n-m}$, где $F_{\alpha, 1, n-m}$ - критическая точка, то гипотеза о незначимости не отклоняется.

Вопрос 8

Доступный обобщенный МНК. В случае обнаружения эффекта гетероскедастичности можно провести коррекцию оценок параметров модели с помощью взвешенного МНК с весами $\hat{\sigma}_t^{-2} = h^{-1}(z_t^T \hat{\alpha})$. Может оказаться, что для некоторых t $h^{-1}(z_t^T \hat{\alpha}) < 0$. Если число таких наблюдений невелико, то их можно просто выбросить, в противном случае использовать представление $\sigma_t^2 = \exp(h(z_t^T \alpha))$.

Вопрос 9

В качестве критерия величины смещения (функции риска) выберем среднюю сумму квадратов ошибок (ССКО)

$$L_{\tilde{\theta}}(\theta) = E(\tilde{\theta} - \theta)^T (\tilde{\theta} - \theta),$$

где $\tilde{\theta}$ - некоторая оценка параметров $\theta \in R^m$.

Для МНК оценки

$$L_{\hat{\theta}}(\theta) = \sigma^2 \text{tr}(X^T X)^{-1} \rightarrow \infty \text{ при } \lambda_1 \rightarrow 0,$$

т.е. МНК оценка в условиях усиления коллинеарности ($\lambda_1 \rightarrow 0$) не обеспечивает ограниченность ССКО. Это будет происходить и в условиях увеличения числа наблюдений. В нормальных условиях (при отсутствии мультиколлинеарности) $\text{tr}(X^T X)^{-1} \rightarrow 0$ при $N \rightarrow \infty$, т.е. теоретически гарантируется состоятельность в среднеквадратичном.

Вопрос 10

Пусть истинное уравнение (4.2), а мы оцениваем регрессию

$$y_u = \gamma_1 \omega_{u1} + \dots + \gamma_k \omega_{uk} + \xi_u \quad (4.5)$$

или

$$y = W\gamma + \xi, \quad X = \begin{bmatrix} W & V \end{bmatrix}$$

Другими словами, $\omega_1 = x_1, \dots, \omega_k = x_k$ входят в регрессию (4.5), а остальные $m - k$ переменных $\omega_{k+1}, \dots, \omega_{m-k} = x_m$ в регрессии отсутствуют.

Докажем, что тогда используемая оценка

$$\hat{\gamma} = (W^T W)^{-1} W^T y \quad (4.6)$$

в общем случае является смещённой в том смысле, что $E\hat{\gamma} \neq \alpha_1$.

Действительно,

$$E\hat{\gamma} = (W^T W)^{-1} W^T E y = (W^T W)^{-1} W^T X \alpha = (W^T W)^{-1} W^T \begin{bmatrix} W & V \end{bmatrix} \cdot \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix},$$

где α_1 и α_2 - подвекторы вектора α размерности k и $m - k$ соответственно.

Перемножая, получим

$$E\hat{\gamma} = \begin{bmatrix} I_k & (W^T W)^{-1} W^T V \end{bmatrix} \cdot \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \alpha_1 + (W^T W)^{-1} W^T V \alpha_2 \quad \blacksquare$$

Оценка $\hat{\gamma}$ будет несмещённой, когда матрицы W и V ортогональны, то есть $W^T V = 0$.

Можно показать также, что МНК оценка в случае недобора является несостоятельной.

Вопрос 11

В таблице исходных данных могут присутствовать наблюдения (например, наблюдение под номером i), сильно влияющие на оценки параметров. Если эти наблюдения связаны с матрицей X , то наблюдение x_i называется влияющим или точкой разбалансировки. Точки разбалансировки не обязательно дают большие остатки в соответствующих наблюдениях, эти большие остатки могут появиться в других наблюдениях.

Вопрос 12

В 4 раза.

Вопрос 13

Перебор регрессоров приводит к неэффективности оценок. Недобор – к смещению.

Вопрос 14

Критерий непротиворечивости, абсолютно помехоустойчивый критерий.

Вопрос 15

Рассмотрим более общий случай, когда среди опытов, входящих в эксперимент, имеются повторные.

Без ограничения общности можно сгруппировать опыты по сериям так чтобы каждая из них содержала все повторные опыты, которые проведены при одних и тех же условиях.

Положим, что всего имеются N_c серий и в каждую v -ю серию входят r_v - повторных опытов, $v = \overline{1, N_c}$. Общее число опытов $N = \sum_{v=1}^{N_c} r_v$

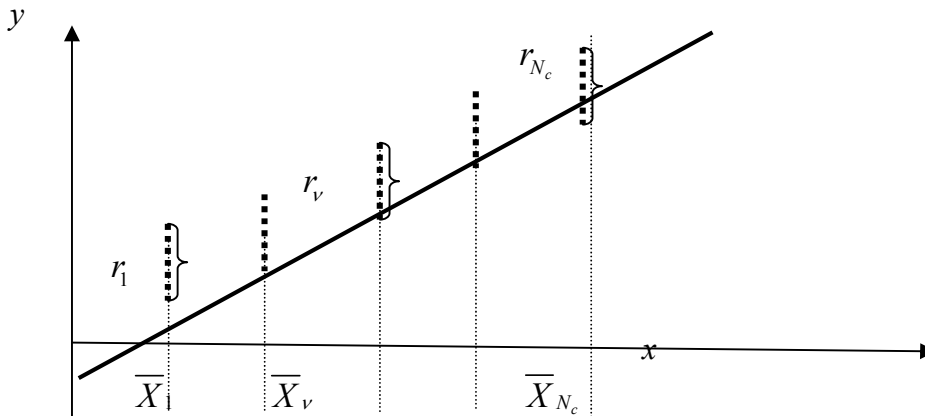
Как и раньше рассматриваем линейную модель наблюдения

$$y(x) = f^T(x)\theta + \varepsilon \quad (2.14)$$

или в матричной форме

$$y = X\theta + \varepsilon. \quad (2.15)$$

Будем считать, что элементы векторов y, ε и строки матрицы X сгруппированы по сериям.



Введем матрицу усреднения

$$S = \begin{pmatrix} r_1^{-1} I_{r_1}^T & & & 0 \\ & r_2^{-1} I_{r_2}^T & & \\ & & \ddots & \\ 0 & & & r_{N_c}^{-1} I_{r_{N_c}}^T \end{pmatrix} \quad \text{размером } N_c \times N_c,$$

где I_{r_v} - вектор столбец размерности r_v , состоящий из единиц. Из структуры этой матрицы видно, что

$$SS^T = R^{-1} = (r_i^{-1} \delta_{ij}), \quad i, j = \overline{1, N_c}, \quad (2.16)$$

где R - диагональная матрица весов, выражающая распределения опытов по сериям.

Умножим теперь левую и правую часть (2.15) слева на матрицу S и получим

$$\bar{y} = \bar{X}\theta + \varepsilon. \quad (2.17)$$

Легко понять, что вектор $\bar{y}$ в качестве элементов имеет среднее значения отклика по сериям:

$$\bar{y} = Sy = (\bar{y}_1, \bar{y}_2, \dots, \bar{y}_{N_c})^T, \quad \text{причем } \bar{y}_v = \bar{y}(x_v) = r_v^{-1} \sum_{j=1}^{r_v} y_{v_j}, \quad v = \overline{1, N_c}.$$

Матрица $\bar{X}$ содержит N_c строк, которые отличаются друг от друга и соответствуют разным сериям:

$$\bar{X} = SX = (f(x_1), f(x_2), \dots, f(x_{N_c}))^T.$$

Вектор усреднённых ошибок имеет вид $\bar{\varepsilon} = (\bar{\varepsilon}_1, \dots, \bar{\varepsilon}_{N_c})^T$, где $\bar{\varepsilon}_v = r_v^{-1} \sum_{j=1}^{r_v} \varepsilon_{v_j}$, $v = \overline{1, N_c}$.

Учитывая, что

$$E(\varepsilon) = O_{N,1}; \quad E(\varepsilon \varepsilon^T) = D(\varepsilon) = I_N \sigma^2,$$

Получим

$$E(\bar{\varepsilon}) = O_{N_c,1}; \quad E(\bar{\varepsilon} \bar{\varepsilon}^T) = D(\bar{\varepsilon}) = R^{-1} \sigma^2 = S I_N \sigma^2 S^T.$$

Сумму взвешенных средних квадратичных отклонений для модели (2.15) можно записать в виде квадратичной формы

$$S(\theta) = \varepsilon^T D^{-1}(\varepsilon) \varepsilon = (y - X\theta)^T \sigma^{-2} (y - X\theta).$$

Опуская сомножитель σ^{-2} , запишем

$$SS(\theta) = (y - X\theta)^T (y - X\theta). \quad (2.18)$$

Для модели (2.17) аналогичная мера рассогласования будет иметь вид

$$S_1(\theta) = \bar{\varepsilon}^T D^{-1}(\bar{\varepsilon}) \bar{\varepsilon} = (\bar{y} - \bar{X}\theta)^T R \sigma^{-2} (\bar{y} - \bar{X}\theta).$$

Опуская σ^{-2} , запишем

$$SS_1(\theta) = (\bar{y} - \bar{X}\theta)^T R (\bar{y} - \bar{X}\theta). \quad (2.19)$$

Матрица весов R может быть представлена в виде произведения

$$R = Q^T Q, \quad (2.20)$$

где Q - матрица размером $N \times N_c$

$$Q = \begin{bmatrix} I_{r_1} & . & . & . & 0 \\ . & I_{r_2} & & & . \\ . & & . & & . \\ . & & & . & . \\ 0 & . & . & . & I_{r_{N_c}} \end{bmatrix}, \quad \text{где } I_{r_v} - \text{столбец единиц}$$

Подставляя (2.20) в (2.19) получим

$$SS_1(\theta) = (Q\bar{y} - X\theta)^T (Q\bar{y} - X\theta), \quad (2.21)$$

где вектор $Q\bar{y}$ – имеет уже N – составляющих, как и вектор y .

Суммы $SS(\theta)$ и $SS_1(\theta)$ связаны между собой. Используя перегруппировку слагаемых первой суммы, можно показать, что

$$SS(\theta) = SS_1(\theta) + SS_E, \quad (2.22)$$

где

$$SS_E = (y - Q\bar{y})^T (y - Q\bar{y}) \quad (2.23)$$

и не зависима от θ . Поэтому совершенно безразлично какой мерой рассогласования пользоваться для отыскания оценок параметров: (2.18), (2.19) или (2.21).

Точечную оценку дисперсии ошибки эксперимента в случае, когда имеются повторные опыты лучше всего определять через сумму (2.23) по формуле

$\hat{\sigma}_E^2 = SS_E / f_E$, где f_E – число степеней свободы суммы SS_E , равное

$f_E = N - N_c = \sum_{v=1}^{N_c} (r_v - 1)$. Оценка $\hat{\sigma}_E^2$ есть не что иное, как объединенная оценка

выборочных дисперсий ошибки опыта, которые можно найти по результатам отдельных серий повторных опытов. Оценка дисперсии ошибки в v -ой серии повторных опытов равна

$$\hat{\sigma}_{Ev}^2 = \frac{SS_{Ev}}{f_{Ev}}, \text{ где } SS_{Ev} = \sum_{j=1}^{r_v} (y_{vj} - \bar{y}_v)^2, \quad f_{Ev} = r_v - 1.$$

Отсюда следует, что объединенная оценка по всем сериям имеет вид

$$\hat{\sigma}_E^2 = \sum_{v=1}^{N_c} \frac{f_{Ev}}{f_E} \hat{\sigma}_{Ev}^2 = \sum_{v=1}^{N_c} SS_{Ev} / f_E = SS_E / f_E.$$

Заметим, что оценка $\hat{\sigma}_E^2$ определяется независимо от модели.

При $\theta = \hat{\theta}$ имеем следующие соотношения:

$$SS_R \equiv SS(\hat{\theta}) = SS_1(\hat{\theta}) + SS_E, \quad (2.24)$$

где SS_R – остаточная сумма со степенями свободы $f_R = N - m$;

$SS_{LF} = SS_1(\hat{\theta})$ – мера степени неадекватности представления экспери-

ментальных данных с помощью выбранной модели. Она оценивает рассеивания средних значений отклика по сериям относительно расчетной регрессионной зависимости с учётом весов (**Lack of fit – неадекватность**).

Оценка дисперсии на основе суммы SS_{LF} вычисляется по формуле

$$\hat{\sigma}_{LF}^2 = SS_{LF} / f_{LF}, \quad f_{LF} = N_c - m. \quad (2.25)$$

Оценка величины σ^2 на основе SS_R выражается отношением

$$\hat{\sigma}_R^2 = SS_R / f_R, \quad f_R = N - m.$$

Оценка величины $\hat{\sigma}_R^2$ является, по существу, объединённой. Используя (2.24) и учитывая, что $f_R = f_{LF} + f_E$ можно записать

$$\hat{\sigma}_R^2 = (f_{LF} / f_R) \hat{\sigma}_{LF}^2 + (f_E / f_R) \hat{\sigma}_E^2.$$

Задача

1. Модель $\eta_1 = \theta_1 x + \theta_2 x^2$
2. Гипотеза принимается

Билет №7

Вопрос 1

Часто класс допустимых решений задают в виде параметрического множества функций.

Вопрос 2

Оценки параметров распределены по нормальному закону с мат.ожиданием равным истинным значениям параметров и дисперсией $\sigma^2(X^T X)^{-1}$

Вопрос 3

$$D(\hat{\theta}) = E(\hat{\theta} - E\hat{\theta})(\hat{\theta} - E\hat{\theta})^T = E[(X^T X)^{-1} X^T \varepsilon \varepsilon^T X (X^T X)^{-1}] = \sigma^2 (X^T X)^{-1} \blacksquare$$

Вопрос 4

Это проверка того, что модель системы не изменилась со временем.

Вопрос 5

Число главных компонент определяется по числу относительно больших собственных значений информационной матрицы

Вопрос 6

$$I - P = I - X(X^T X)^{-1} X^T.$$

$$(I - P)^2 = I - 2P + P^2 =$$

$$= I - 2X(X^T X)^{-1} X^T + X(X^T X)^{-1} X^T X (X^T X)^{-1} X^T = I - P$$

Вопрос 7

Для линейной модели с двумя факторами при 4 наблюдениях в углах квадрата информационная матрица будет диагональной с диагональю 4. При 8 наблюдениях (повторы) диагональ будет равна 8 и т.д.

Вопрос 8

Доступный обобщенный МНК. В случае обнаружения эффекта гетероскедастичности можно провести коррекцию оценок параметров модели с помощью взвешенного МНК с весами $\hat{\sigma}_t^{-2} = h^{-1}(z_t^T \hat{\alpha})$. Может оказаться, что для некоторых t $h^{-1}(z_t^T \hat{\alpha}) < 0$. Если число таких наблюдений невелико, то их можно просто выбросить, в противном случае использовать представление $\sigma_t^2 = \exp(h(z_t^T \alpha))$.

Вопрос 9

В качестве критерия величины смещения (функции риска) выберем среднюю сумму квадратов ошибок (ССКО)

$$L_{\tilde{\theta}}(\theta) = E(\tilde{\theta} - \theta)^T (\tilde{\theta} - \theta),$$

где $\tilde{\theta}$ - некоторая оценка параметров $\theta \in R^m$.

Для МНК оценки

$$L_{\hat{\theta}}(\theta) = \sigma^2 \text{tr}(X^T X)^{-1} \rightarrow \infty \text{ при } \lambda_1 \rightarrow 0,$$

т.е. МНК оценка в условиях усиления коллинеарности ($\lambda_1 \rightarrow 0$) не обеспечивает ограниченность ССКО. Это будет происходить и в условиях увеличения числа наблюдений. В нормальных условиях (при отсутствии мультиколлинеарности) $\text{tr}(X^T X)^{-1} \rightarrow 0$ при $N \rightarrow \infty$, т.е. теоретически гарантируется состоятельность в среднеквадратичном.

Вопрос 10

Переменные, которым соответствуют элементы вектора $\underline{y}$ для минимального собственного числа с значимо отличающимися от 0 значениями и порождают мультиколлинеарность.

Вопрос 11

$$p_{ii} = f^T(x_i)(X^T X)^{-1} f(x_i)$$

. На практике руководствуются следующими правилами. Значения $p_{ii} \leq 0,2$ позволяют считать надежными данные наблюдения, значения $0,2 \leq p_{ii} \leq 0,5$ относят данное наблюдение к классу рискованных, а при значениях $p_{ii} > 0,5$ i -е наблюдение можно принять за точку разбалансировки.

Вопрос 12

В 4 раза.

Вопрос 13

Монотонность коэффициента детерминации от числа включенных в модель регрессоров легко доказывается, если обратиться к его выражению:

$$R^2 = 1 - \frac{RSS}{\sum (y_i - \bar{y})^2}$$

Вопрос 14

При наличии повторной выборки можно предложить критерий селекции, в основу которого положена идея идеального внешнего критерия. При вычислении идеального внешнего критерия вместо зашумленных значений отклика используются точные (незашумленные) значения. Однако в реальной ситуации эти значения нам не известны. В случае повторной выборки вместо неизвестных значений вектора можно использовать и их оценку в виде выборочного среднего $\bar{y}$:

$$\bar{y}^T = (\bar{y}_A^T, \bar{y}_B^T), \bar{y}_{A_i} = (y_{A_i} + y_{B_i})/2, i = 1, \dots, n_A, \bar{y}_B = \bar{y}_A.$$

С использованием вектора $\bar{y}$ можно предложить следующий субидеальный критерий стабильности:

$$\tilde{S}^2 = \|\bar{y}_A - X_A \hat{\theta}_A\|^2 + \|\bar{y}_B - X_B \hat{\theta}_B\|^2.$$

Вопрос 15

С целью управления масштабом оценок введем в рассмотрение функцию стоимости

$$C = \sum_{i=1}^n (y_i - f(x_i)^T \theta)^2 + \sum_{j=1}^m \lambda_j \theta_j^2,$$

где второе слагаемое рассматривается как штраф при условии, что $\lambda_j \geq 0$. Решим данную оптимизационную задачу $C \rightarrow \min$.

Представим целевую функцию в матричном виде

$$C = (y - X\theta)^T (y - X\theta) + \theta^T \Lambda \theta,$$

$$\text{где } \Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_m \end{pmatrix}.$$

Найдем решение из условия равенства нулю первых производных:

$$\frac{\partial C}{\partial \theta_j} = 2X^T y + 2X^T X \hat{\theta} + 2\Lambda \hat{\theta} = 0,$$

$$\hat{\theta} = (X^T X + \Lambda)^{-1} X^T y.$$

Ридж-оценки впервые были введены А. Гоэрлом при исследовании регрессии 2-го порядка (ridge – гребень, хребет). Название "ридж-оценки" навеяно следующими наблюдениями. При сильной мультиколлинеарности линии равного уровня доверительного эллипсоида для разного уровня значимости (доверительной вероятности) становятся сильно вытянутыми и визуально похожими на хребет на географической карте. Данные оценки использовались именно в условиях мультиколлинеарности и поэтому получили название ридж-оценок.

Понятно, что ридж-оценки существенно зависят от выбранной матрицы Λ , в силу чего имеют обозначение

$$\hat{\theta}(\Lambda) = (X^T X + \Lambda)^{-1} X^T y.$$

Часто матрицу Λ задают диагональной в виде $\Lambda_{ii} = \lambda \cdot (X^T X)_{ii}$, $\lambda \geq 0$, т.е. пропорциональной диагональным элементам матрицы $X^T X$. Иногда применяют ещё более простой вариант $\Lambda = \lambda \cdot I_m$. Можно показать, что к данному случаю сводится предыдущий.

Обозначим

$$D = \begin{pmatrix} (X^T X)_{11} & & 0 \\ & \ddots & \\ 0 & & (X^T X)_{mm} \end{pmatrix},$$

$$Z = XD^{-1/2}, \quad \beta = D^{1/2} \theta.$$

Тогда исходная модель

$$y = X\theta + \varepsilon$$

сводится в новых обозначениях к модели

$$y = Z\beta + \varepsilon = XD^{-1/2} D^{1/2} \theta + \varepsilon = X\theta + \varepsilon.$$

Ридж-оценка $\hat{\beta}(\lambda)$ с матрицей $\lambda \cdot I_m$ будет равна

$$\hat{\beta}(\lambda) = (Z^T Z + \lambda I)^{-1} Z^T y = (D^{-1/2} X^T X D^{-1/2} + \lambda I)^{-1} D^{-1/2} X^T y =$$

$$= (X^T X + \lambda D)^{-1} D D^{-1/2} X^T y = D^{1/2} (X^T X + \Lambda)^{-1} X^T y,$$

где $\Lambda_{ii} = \chi X^T X)_{ii}$. Поэтому $\hat{\theta}(\lambda) = D^{-1/2} \hat{\beta}(\lambda)$.

Итак, можно рассматривать более простой случай

$$\hat{\beta}(\lambda) = (Z^T Z + \lambda I)^{-1} Z^T y.$$

Свойства этой оценки:

- 1) $\hat{\beta}(0)$ - МНК-оценка $\hat{\beta}$;
- 2) $\hat{\beta}(\lambda)$ - является линейным преобразованием МНК-оценки и является смещённой;
Действительно,

$$Z^T y = (I + \lambda(Z^T Z)^{-1})^{-1} (Z^T Z)^{-1} Z^T y = B \hat{\beta},$$

где $B = (I + \lambda(Z^T Z)^{-1})^{-1}$ - матрица линейного преобразования и $B \neq I$, что приводит к смещению оценок $\hat{\beta}(\lambda)$ в отличие от $\hat{\beta}$, которые несмещены.

- 3) в классе оценок с фиксированной длиной ридж-оценка минимизирует сумму квадратов отклонений;
- 4) верно и обратное утверждение: ридж-оценка имеет минимальную длину в классе оценок с заданным значением суммы квадратов отклонений (RSS).

Доказано, что всегда существует такое значение параметра λ , что ССКО для оценки $\hat{\beta}(\lambda)$ будет меньше чем ССКО для оценки $\hat{\beta}(0)$. Показано также, что

$$\hat{\beta}^T(\lambda) \hat{\beta}(\lambda) = (\mu_{\max} / (\mu_{\max} + \lambda)) \hat{\beta}^T \hat{\beta},$$

где $\mu_{\max}$ - максимальное собственное число информационной матрицы $Z^T Z$. Это означает, что длина ридж-оценки будет уменьшаться с увеличением параметра регуляризации, но это и так понятно из рассмотрения целевой функции, где используется штраф за длину оценки.

Задача

1. Модель $\eta_1 = \theta_1 x + \theta_2 x^2$
2. Гипотеза принимается

Билет № 8

Вопрос 1

Нужно отметить, что использование симметричных форм критериев всегда целесообразно: во-первых, снижается чувствительность критерия к способу разделения данных на части, и, во-вторых, усредняется влияние помех, действующих на различных частях выборки.

Вопрос 2

Для того, что бы было возможным обращение информационной матрицы.

Вопрос 3

Для получения несмещенной оценки.

Вопрос 4

Ридж-оценки впервые были введены А. Гоэрлом при исследовании регрессии 2-го порядка (ridge – гребень, хребет). Название "ридж-оценки" навеяно следующими наблюдениями. При сильной мультиколлинеарности линии равного уровня доверительного эллипсоида для разного уровня значимости (доверительной вероятности) становятся сильно вытянутыми и визуально похожими на хребет на географической карте. Данные оценки использовались именно в условиях мультиколлинеарности и поэтому получили название ридж-оценок.

Вопрос 5

Необходимо промасштабировать информационную матрицу на ее след.

Вопрос 6

Условие минимума известно – оно определяется как

$$X^T \underbrace{X \hat{\theta}}_{\hat{y}} = X^T y.$$

Отсюда $X^T (y - X \hat{\theta}) = 0$ и далее $X^T (y - \hat{y}) = 0$, т.е. вектор остатков $\hat{e} = y - \hat{y}$ ортогонален Ω .

Вопрос 7

В 2 раза.

Вопрос 8

Ясно, что в реальных ситуациях редко бывает так, что параметр автокорреляции ρ известен. При неизвестном ρ процедуры оценивания параметров θ носят итеративный характер.

Процедура Кохрейна-Оркатта.

6. Для исходной модели оцениваются $\hat{\theta}$ и $\hat{e}$;
7. Рассматривается регрессия $\hat{e}_t = \rho \hat{e}_{t-1} + v_t$ и вычисляется оценка $\hat{\rho}$;
8. Проводится ОМНК оценивание исходной модели с $\hat{\Omega}^{-1}$, где в качестве ρ используется его оценка $\hat{\rho}$;
9. Вычисляется новый вектор остатков $\hat{e}$;
10. Процедура повторяется с п.2.

Останов производится после нескольких итераций или после того как очередное $\hat{\rho}$ мало отличается от предыдущего.

Процедура Хилдрета-Лу.

По схеме ОМНК производится решение следующей экстремальной задачи:

$$\hat{\rho} = \text{Arg} \min_{\rho \in [-1, +1]} \text{RSS}.$$

Попутно при этом мы получаем оценки для вектора параметров θ .

Вопрос 9

$$\hat{\theta} = (X^T X)^{-1} X^T y. \quad \hat{\theta} = (n)^{-1} X^T y = \sum_{i=1}^n y_i / n$$

Вопрос 10

$$P = X(X^T X)^{-1} X^T. \quad P^2 = X(X^T X)^{-1} X^T X(X^T X(X^T X)^{-1} X^T)^{-1} X^T = P$$

Вопрос 11

Точечную оценку дисперсии ошибки эксперимента в случае, когда имеются повторные опыты лучше всего определять через сумму (2.23) по формуле

$\hat{\sigma}_E^2 = SS_E / f_E$, где f_E – число степеней свободы суммы SS_E , равное

$f_E = N - N_c = \sum_{v=1}^{N_c} (r_v - 1)$. Оценка $\hat{\sigma}_E^2$ есть не что иное, как объединенная оценка

выборочных дисперсий ошибки опыта, которые можно найти по результатам отдельных серий повторных опытов. Оценка дисперсии ошибки в v -ой серии повторных опытов равна

$$\hat{\sigma}_{Ev}^2 = \frac{SS_{Ev}}{f_{Ev}}, \text{ где } SS_{Ev} = \sum_{j=1}^{r_v} (y_{vj} - \bar{y}_v)^2, \quad f_{Ev} = r_v - 1.$$

Отсюда следует, что объединенная оценка по всем сериям имеет вид

$$\hat{\sigma}_E^2 = \sum_{v=1}^{N_c} \frac{f_{Ev}}{f_E} \hat{\sigma}_{Ev}^2 = \sum_{v=1}^{N_c} SS_{Ev} / f_E = SS_E / f_E.$$

Вопрос 12

. Рассмотрим случай когда в модели присутствует свободный член. В этом случае сумма остатков равно нулю, т.е.

$$\sum (y_i - \hat{y}_i) = 0. \quad (a)$$

Покажем это. Для этого сумму квадратов отклонений для модели со свободным членом

$$\sum (y_i - \theta_0 - \theta_1 x_{1i} - \dots - \theta_{m-1} x_{m-1,i})^2$$

продифференцируем по θ_0 :

$$\sum (y_i - \hat{\theta}_0 - \hat{\theta}_1 x_{1i} - \dots - \hat{\theta}_{m-1} x_{m-1,i}) = 0 \text{ или } \sum (y_i - \hat{y}_i) = 0.$$

Вопрос 13

Для ее вычисления необходимо обратить матрицу коэффициентов корреляций, которая может быть плохо обусловлена в условиях мультиколлинеарности.

Вопрос 14

$$p_{ii} = f^T(x_i)(X^T X)^{-1} f(x_i)$$

. На практике руководствуются следующими правилами. Значения $p_{ii} \leq 0,2$ позволяют считать надежными данные наблюдения, значения $0,2 \leq p_{ii} \leq 0,5$ относят данное наблюдение к классу рискованных, а при значениях $p_{ii} > 0,5$ i -е наблюдение можно принять за точку разбалансировки.

Вопрос 15

Заменим предположение A.2 на более общее предположение

$$A.5 \quad E(\varepsilon \varepsilon^T) = \sigma^2 \Omega = V,$$

где $tr \Omega = n$, Ω - симметричная п.о.

Эффективный оценщик для θ в этом случае можно получить, если исходную модель трансформировать в модель, соответствующую условиям теоремы Гаусса – Маркова.

Поскольку Ω - симметричная п.о., то её можно выразить как PP^T , где P – невырожденная матрица, т.е.

$$\Omega = PP^T \text{ и } \Omega^{-1} = (P^T)^{-1}P^{-1} = (P^{-1})^T P^{-1}.$$

(Свойства обратной матрицы $(AB)^{-1} = B^{-1}A^{-1}$).

Перепишем модель (1.4):

$$y^* = X^* \theta + \varepsilon^*,$$

$$\text{где } y^* = P^{-1}y, \quad X^* = P^{-1}X, \quad \varepsilon^* = P^{-1}\varepsilon.$$

Для данной модели выполняются свойства:

$$E(\varepsilon^*) = P^{-1}E(\varepsilon) = 0, \quad E(\varepsilon^* \varepsilon^{*T}) = \sigma^2 I, \quad \text{rg} X^* = \text{rg} P^{-1}X = \text{rg} X = m.$$

Мы можем теперь формально применить теорему Гаусса-Маркова для вычисления оценок:

$$\begin{aligned} \hat{\theta}_{\text{ОМНК}} &= (X^{*T} X^*)^{-1} X^{*T} y^* = (X^T P^{-1T} P^{-1} X)^{-1} X^T P^{-1T} P^{-1} y = \\ &= (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} y = (X^T V^{-1} X)^{-1} X^T V^{-1} y. \end{aligned} \quad (3.1)$$

Данные оценки получили название оценок по обобщенному МНК.

Их также можно получить отталкиваясь от нормированной на дисперсионную матрицу сумму квадратов ошибок:

$$\begin{aligned} S(\theta) &= (y - X\theta)^T V^{-1} (y - X\theta) = \\ &= y^T V^{-1} y - 2\theta^T X^T V^{-1} y + \theta^T X^T V^{-1} X \theta. \end{aligned}$$

Первые частные производные $S(\theta)$ по параметрам θ имеют вид

$$\frac{\partial S(\theta)}{\partial \theta} = -2X^T V^{-1} y + 2X^T V^{-1} X \theta.$$

Приравнявая их нулю, получаем систему нормальных уравнений

$$X^T V^{-1} X \hat{\theta}_{\text{ОМНК}} = X^T V^{-1} y \quad \text{с решением } \hat{\theta}_{\text{ОМНК}} = (X^T V^{-1} X)^{-1} X^T V^{-1} y.$$

Дисперсионная матрица оценок имеет вид

$$D(\hat{\theta}_{\text{ОМНК}}) = \sigma^2 (X^{*T} X^*)^{-1} = \sigma^2 (X^T \Omega^{-1} X)^{-1} = (X^T V^{-1} X)^{-1}.$$

Если в структуре матрицы $E(\varepsilon \varepsilon^T) = \sigma^2 \Omega = V$ неизвестной величиной является дисперсия σ^2 , то ее несмещенной оценкой будет оценка:

$$\hat{\sigma}^2 = \frac{1}{n-m} (Y - X \hat{\theta}_{\text{ОМНК}})^T \Omega^{-1} (Y - X \hat{\theta}_{\text{ОМНК}}).$$

Задача

-
1. Модель $\eta_1 = \theta_1 + \theta_2 x$
 2. Гипотеза принимается

Билет 9

Вопрос 1

Термин "сферичность" возмущений происходит из того наблюдения, что в многомерном пространстве распределения помехи ε при $Var(\varepsilon) = \sigma^2 I_n$ (случай равноточных наблюдений) точки с одинаковой вероятностью образуют сферы. В рассматриваемом случае неравноточных наблюдений вместо сфер будут образовываться многомерные эллипсоиды.

Вопрос 2

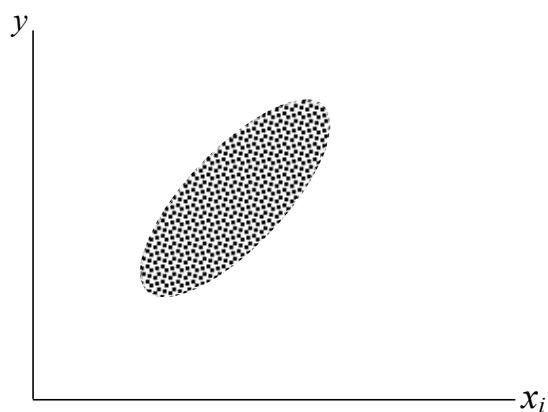
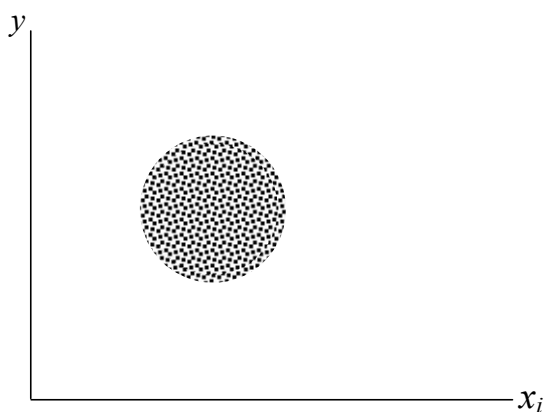
Нет.

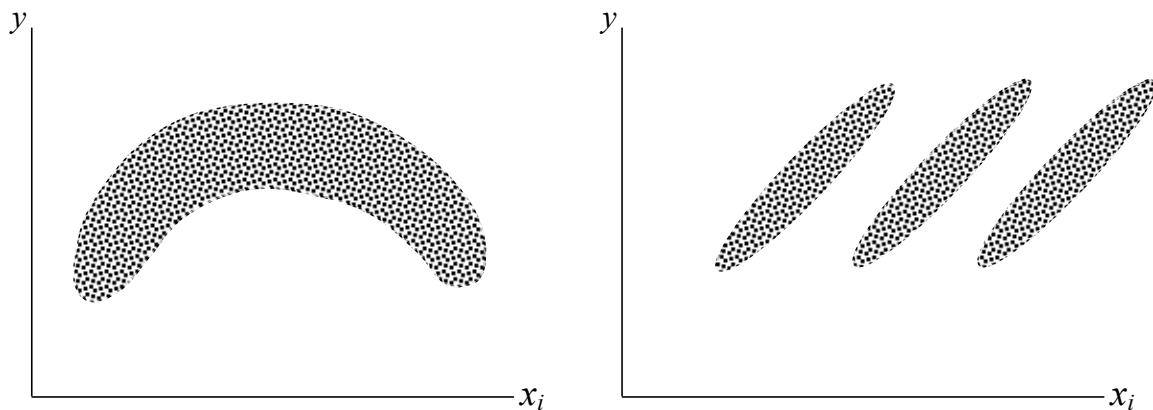
Вопрос 3

При предварительном анализе характера исследуемых зависимостей ограничиваются в основном изучением "геометрии" парных корреляционных полей. Под корреляционным полем переменных (u, v) понимается графическое представление имеющихся измерений $(u_1, v_1), (u_2, v_2), \dots, (u_n, v_n)$ этих переменных в плоскости (u, v) .

Анализ парных корреляционных полей состоит обычно в следующем:

- а) построение на основании имеющихся исходных данных корреляционных полей для всевозможных пар переменных вида (x_j, x_i) и (x_j, y) — всего таких пар будет $k(k+1)/2$. Этот процесс легко автоматизируется с помощью средств современных ЭВМ;
- б) визуальное прослеживание характера вытянутости каждого корреляционного поля (эллипсоидально-линейное, нелинейное-монотонное, с наличием экстремумов и т. п.);





Это примеры корреляционных полей для пассивного эксперимента.

Вопрос 4

Выделяющиеся точки графика: некоторые из остатков могут по абсолютной величине сильно превосходить все остальные остатки. Если максимальное значение остатка $|d_{\max}| > \delta$, где δ - некоторое заранее выбранное число, то наблюдение, имеющее такой остаток, является аномальным. Величина остатка аномального наблюдения зависит от объема выборки n : при $n > 20$ его можно обнаружить по правилу «трех сигм» ($\delta = 3$), при $n < 20$, например, по критерию Смирнова - Граббса. В этом случае вычисляемая статистика имеет вид

$$D = (d_{\max} - \bar{d}) / S_d, \text{ где } \bar{d} - \text{среднее значение остатков, } S_d^2 = 1/n \left(\sum_{i=1}^n (d_i - \bar{d})^2 \right).$$

Вопрос 5

Свойство 4, поскольку преследуется цель уменьшения нормы оценок параметров.

Вопрос 6

Доступный обобщенный МНК. В случае обнаружения эффекта гетероскедастичности можно провести коррекцию оценок параметров модели с помощью взвешенного МНК с весами $\hat{\sigma}_i^{-2} = h^{-1}(z_i^T \hat{\alpha})$. Может оказаться, что для некоторых t $h^{-1}(z_i^T \hat{\alpha}) < 0$. Если число таких наблюдений невелико, то их можно просто выбросить, в противном случае использовать представление $\sigma_i^2 = \exp(h(z_i^T \alpha))$.

Вопрос 7

Для критерия Δ_{abs}^2 , например, его среднее значение будет определяться величиной

$$\bar{\Delta}_{abs}^2 = u^T D_{\Delta_{abs}(p)} u + \sigma^2 \cdot p.$$

С увеличением σ^2 минимум критерия в его зависимости от сложности модели смещается влево, т.е. выбираются все более простые модели. Абсолютно помехоустойчивый критерий получил свое название по причине того, что второе слагаемое для него менее всего возрастает при увеличении дисперсии помехи. Другими словами, при увеличении дисперсии помехи минимум данного критерия с наименьшей скоростью смещается влево в сторону все более простых моделей.

Вопрос 8

Алгоритм состоит в следующем.

На первом шаге выбирается некоторое число лучших моделей, состоящих из одного регрессора. На втором этапе к этим моделям добавляются различные регрессоры по одному, и отбираются модели с двумя аргументами, улучшающие величину критерия, и т.д. до его ухудшения или до построения полной модели.

На последнем шаге можно из оставшихся лучших моделей выбрать одну по другому критерию.

Вопрос 9

$$D(\hat{\theta}) = E(\hat{\theta} - E\hat{\theta})(\hat{\theta} - E\hat{\theta})^T = E[(X^T X)^{-1} X^T \varepsilon \varepsilon^T X (X^T X)^{-1}] = \sigma^2 (X^T X)^{-1} \blacksquare$$

Вопрос 10

Имеем $X^T(y - X\hat{\theta}) = 0$ и далее $X^T(y - \hat{y}) = 0$, т.е. вектор остатков $\hat{e} = y - \hat{y}$ ортогонален Ω . В тоже время справедливо представление $y = \hat{y} + \hat{e}$. Становится понятным, что $\hat{y}$ есть проекция y на $R(X)$, а именно $\hat{y} = Py$. Можно рассматривать также подпространство

$R^\perp(X)$ – ортогональное дополнение для $R(X)$ в $R(X)$. Поскольку

$$\hat{e} = y - \hat{y} = y - X\hat{\theta} = (I_n - X(X^T X)^{-1} X^T)y = (I_n - P)y,$$

то $M = I_n - P$ – есть матрица проектирования вектора y на подпространство $R^\perp(X)$.

Вопрос 11

$$\hat{\theta} = (X^T X)^{-1} X^T y. \quad \hat{\theta} = (n)^{-1} X^T y = \sum_{i=1}^n y_i / n$$

Вопрос 12

В качестве критерия величины смещения (функции риска) выберем среднюю сумму квадратов ошибок (ССКО)

$$L_{\tilde{\theta}}(\theta) = E(\tilde{\theta} - \theta)^T (\tilde{\theta} - \theta),$$

где $\tilde{\theta}$ – некоторая оценка параметров $\theta \in R^m$.

Для МНК оценки

$$L_{\hat{\theta}}(\theta) = \sigma^2 \text{tr}(X^T X)^{-1} \rightarrow \infty \text{ при } \lambda_1 \rightarrow 0,$$

т.е. МНК оценка в условиях усиления коллинеарности ($\lambda_1 \rightarrow 0$) не обеспечивает ограниченность ССКО. Это будет происходить и в условиях увеличения числа наблюдений. В нормальных условиях (при отсутствии мультиколлинеарности) $\text{tr}(X^T X)^{-1} \rightarrow 0$ при $N \rightarrow \infty$, т.е. теоретически гарантируется состоятельность в среднеквадратичном.

Вопрос 13

Переменные, которым соответствуют элементы вектора y для минимального собственного числа с значимо отличающимися от 0 значениями и порождают мультиколлинеарность.

Вопрос 14

Статистика всегда удовлетворяет неравенству

$$0 < DW < 4.$$

При отсутствии автокорреляции статистика DW близка к двум, при положительной автокорреляции она близка к нулю, а при отрицательной к 4. Точное распределение статистики DW зависит не только от числа наблюдений n и количества регрессоров m , но и от всей матрицы X , что означает невозможность ее табулирования. Тем не менее, Дарбин и Утсон показали, что есть определенные границы, зависящие только от числа наблюдений n , количества регрессоров m и уровня значимости (поэтому могут быть затабулированы).

Нулевая гипотеза состоит в том, что $H_0: \rho = 0$, против альтернативы $H_1: \rho > 0$ или

$H_1: \rho < 0$. Обозначим и d_α^U , d_α^L – нижний и верхний уровни для доверительной границы,

которые не зависят от X . Влияние неопределенности в матрице X все же привело к появлению зоны неопределенности в принятии или отвержении первоначальной гипотезы. Результаты можно представить

$4-d_{\alpha}^L < DW < 4$	гипотеза $H_0: \rho = 0$ отвергается, есть отрицательная корреляция;
$4-d_{\alpha}^U < DW < 4-d_{\alpha}^L$	неопределенность;
$d_{\alpha}^U < DW < 4-d_{\alpha}^U$	$H_0: \rho = 0$ не отвергается;
$d_{\alpha}^L < DW < d_{\alpha}^U$	неопределенность;
$0 < DW < d_{\alpha}^L$	$H_0: \rho = 0$ отвергается, есть положительная корреляция.

Вопрос 15

Модифицированные жордановы исключения или выметания

Говорят, что матрица $A = [(a_{ij})]$, $i, j = \overline{1, m}$ выметена по k -ой строке и k -му столбцу (или по k -му ведущему элементу), если она преобразована к матрице $A^* = [(a_{ij}^*)]$, у которой

$$a_{kk}^* = \frac{1}{a_{kk}}, \quad a_{ik}^* = \frac{-a_{ik}}{a_{kk}} \quad (i \neq k),$$

$$a_{kj}^* = \frac{a_{kj}}{a_{kk}} \quad (j \neq k), \quad a_{ij}^* = a_{ij} - \frac{a_{ik}a_{kj}}{a_{kk}} \quad (i, j \neq k).$$

Оператор выметания обладает следующими свойствами:

- 3) оператор выметания обратим, так что двукратное применение выметания по одному и тому же (в смысле номера) ведущему элементу матрицы равносильно неприменению этого оператора.
- 4) Выметание коммутативно, так что выметание в порядке k, l -ых элементов равносильно выметанию, проведенному в порядке l, k .

Сформируем матрицу A следующего вида

$$A = \begin{matrix} & \begin{matrix} \overset{m}{X^T X} & \overset{m+1}{X^T y} \end{matrix} \\ \begin{matrix} \overset{m}{y^T X} \\ \overset{m+1}{y^T y} \end{matrix} & \left| \begin{matrix} \hline \hline \end{matrix} \right. \end{matrix} \quad (4.43)$$

Если матрицу A "выметим" по ее первым m ведущим элементам, то на ее месте получим (при условии, что $(X^T X)^{-1}$ существует)

$$B = \left(\begin{array}{c|c} (X^T X)^{-1} & (X^T X)^{-1} X^T y \\ \hline -y^T X (X^T X)^{-1} & y^T y - y^T X (X^T X)^{-1} X^T y \end{array} \right) = \left(\begin{array}{cc} (X^T X)^{-1} & \hat{\theta} \\ -\hat{\theta}^T & RSS \end{array} \right).$$

Если провести выметание по первым i ведущим элементам ($i \leq m$), то в $m+1$ -ом столбце на первых i позициях будут находиться $\hat{\theta}_j$, $j=1, 2, \dots, i$, $i \times i$ -матрица будет

соответствовать $(X_{(i)}^T X_{(i)})^{-1}$ – главному минору. Кроме того, в силу обратимости и коммутативности выметание выполнение выметания по некоторому ведущему элементу матрицы A приводит либо к включению в модель соответствующего регрессора, либо к исключению этого регрессора из модели, если он в ней уже присутствовал.

Порядок построения моделей.

Используя операцию выметания можно построить эффективную процедуру построения всех возможных моделей.

Рассмотрим простой пример. Пусть

$$f(x, \theta) = \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3, \quad m = 3.$$

Всего возможно $2^3 - 1 = 7$ моделей регрессии

№	№ выметаемого элемента	Модель
1	1	$y_1 = \theta_1 x_1$
2	2	$y_2 = \theta_1 x_1 + \theta_2 x_2$
3	1	$y_3 = \theta_2 x_2$
4	3	$y_4 = \theta_2 x_2 + \theta_3 x_3$
5	1	$y_5 = \theta_2 x_2 + \theta_3 x_3 + \theta_1 x_1$
6	2	$y_6 = \theta_3 x_3 + \theta_1 x_1$
7	1	$y_7 = \theta_3 x_3$

Если $m=4$, то оптимальная последовательность будет следующей

$$S_4 = 1 \ 2 \ 1 \ 3 \ 1 \ 2 \ 1 \ 4 \ 1 \ 2 \ 1 \ 3 \ 1 \ 2 \ 1.$$

Применяя метод индукции, получим

$$S_m = S_{m-1} \quad m \quad S_{m-1}.$$

Более экономных последовательностей не существует, поскольку здесь каждое выметание порождает новую модель. Всего нужно произвести $2^m - 1$ выметания, чтобы последовательно построить $2^m - 1$ модель.

Последовательность S_m по имени автора получила название счетчик Гарсайда.

Задача

1. Модель $\eta_1 = \theta_1 + \theta_2 x$
2. Гипотеза отвергается

Тун 1. Установление самого факта наличия (или отсутствия) статистически значимой связи между откликом и факторами.

Выбор формы связи (т.е. класса допустимых решений F и конкретного вида $\bar{f}(\bar{x}) \in F$) и состава предикторов играет второстепенную роль.

Тун 2. Прогноз (восстановление) неизвестных значений интересующих нас результирующих показателей по заданным значениям факторов x .

Тун 3. Выявление причинных связей между откликами и факторами, частичное управление значениями системы путем регулирования величин входных факторов.

Здесь на первый план выходят задача правильного определения структуры модели.

Вопрос 2

Этап 4 (определение класса допустимых решений).

Главной целью является определение общего вида, структуры искомой связи между $\bar{y}$ и $\bar{x}$, или, другими словами, описание класса функций F , в рамках которого будет производиться поиск конкретного вида интересующей зависимости.

Чаще всего это описание дается в форме некоторого параметрического семейства функций $\bar{f}(\bar{x}, \bar{\theta})$, поэтому этап этот называют также этапом параметризации модели.

Вопрос 3.

Линейные оценки представимы в виде

$$\hat{\theta} = Ty,$$

где $y^T = (y_1, \dots, y_n)$, T – матрица $m \times n$.

Вопрос 4

Достоинство – свободно от влияния масштаба. Недосток – выявляет только парную коллинеарность.

Вопрос 5

Определитель информационной матрицы, минимальное собственное число, мера обусловленности по Нейману-Голдстейну, максимальная парная сопряженность, максимальная сопряженность.

Вопрос 6

Число главных компонент и определяется по числу относительно больших собственных значений матрицы $X^{*T}X^*$. Недобор суммарной изменчивости может при этом составлять несколько процентов.

Вопрос 7

Критерий называется внешним, если его определение основано на использовании новой информации, «свежих» наблюдениях, не использованных при синтезе модели.

Вопрос 8.

Возможно использование критерия скользящего контроля и критерия регулярности.

Вопрос 9

$$D(\hat{\theta}) = E(\hat{\theta} - E\hat{\theta})(\hat{\theta} - E\hat{\theta})^T = E[(X^T X)^{-1} X^T \varepsilon \varepsilon^T X (X^T X)^{-1}] = \sigma^2 (X^T X)^{-1} \blacksquare$$

Вопрос 10

Для получения ММП оценок для θ и σ^2 необходимо совместно решить

$$\frac{\partial \log(L)}{\partial \Theta} = -\frac{1}{\sigma^2} X^T (y - X\Theta) = 0$$

$$\frac{\partial \log(L)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (y - X\Theta)^T (y - X\Theta) = 0$$

Решения:

$$\hat{\Theta}_{MML} = (X^T X)^{-1} X^T y, \quad \hat{\sigma}_{MML}^2 = (y - X\hat{\Theta}_{MML})^T (y - X\hat{\Theta}_{MML}) / n.$$

Вопрос 11

Наиболее известный критерий для выявления автокорреляции 1-го порядка – это критерий Дарбина-Уотсона. Для регрессии

$y = X\theta + \varepsilon$ вычисляются остатки $\hat{e} = y - X\hat{\theta}$. По ним вычисляется статистика этого критерия:

$$DW = \sum_{t=2}^n (\hat{e}_t - \hat{e}_{t-1})^2 / \sum_{t=1}^n \hat{e}_t^2.$$

Показано, что

$$DW \cong 2(1 - \hat{\rho}),$$

где $\hat{\rho}$ – коэффициент автокорреляции в регрессии $\hat{e}_t = \rho \hat{e}_{t-1} + v$:

$$\hat{\rho} = \sum_{t=2}^n \hat{e}_t \hat{e}_{t-1} / \sum_{t=2}^n \hat{e}_t^2.$$

Статистика всегда удовлетворяет неравенству

$$0 < DW < 4.$$

При отсутствии автокорреляции статистика DW близка к двум, при положительной автокорреляции она близка к нулю, а при отрицательной к 4.

Вопрос 12

Частный случай 1: проверка равенства отдельного параметра заданной величине c .

Н: $\Theta_j = c$ или Н: $a^T \Theta = c$, где a^T – вектор-строка, в которой на j -ом месте стоит 1, на остальных местах – нули. При $c = 0$ проверяется гипотеза о незначимости параметра.

Обозначим $d_{jj} = ((X^T X)^{-1})_{jj}$ – j -ый диагональный элемент. Тогда

$$F = \frac{(a^T \hat{\Theta} - c)^T (a^T \hat{\Theta} - c)}{\hat{\sigma}^2 d_{jj}} = \frac{(\hat{\Theta}_j - c)^2}{\hat{\sigma}^2 d_{jj}}$$

имеет при гипотезе Н распределение $F_{1, n-m}$. т.е если $F \leq F_{\alpha, 1, n-m}$, где $F_{\alpha, 1, n-m}$ – критическая точка, то гипотеза о незначимости не отклоняется.

Вопрос 13

Главные компоненты должны быть ортогональными, суммарная их изменчивость должна быть равна таковой для исходных факторов, сами они определяются как линейные комбинации исходных факторов.

Вопрос 14

Оператор выметания обладает следующими свойствами:

- 5) оператор выметания обратим, так что двукратное применение выметания по одному и тому же (в смысле номера) ведущему элементу матрицы равносильно неприменению этого оператора.
- 6) Выметание коммутативно, так что выметание в порядке k , l -ых элементов равносильно выметанию, проведенному в порядке l , k .

Если $m=4$, то оптимальная последовательность будет (счетчик Гарсайда) следующей

$$S_4 = 1 \ 2 \ 1 \ 3 \ 1 \ 2 \ 1 \ 4 \ 1 \ 2 \ 1 \ 3 \ 1 \ 2 \ 1.$$

Применяя метод индукции, получим

$$S_m = S_{m-1} \ m \ S_{m-1}.$$

Вопрос 15

Мэлоуэс (Mallows) предложил использовать в качестве оценки для Δ_p статистику

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} + 2p - n,$$

где $\hat{\sigma}^2$ - подходящая оценка для σ^2 , $RSS_p = (y - \hat{y}_p)^T (y - \hat{y}_p)$ - остаточная сумма квадратов.

Часто берут $\hat{\sigma}^2 = \frac{RSS}{(n-m)}$, где m - полное число регрессоров.

Полагая $\hat{\sigma}^2 \approx \sigma^2$, имеем

$$E[C_p] \cong \frac{1}{\sigma^2} E[RSS_p] + 2p - n,$$

далее

$$\begin{aligned} E[RSS_p] &= E[(y - \hat{y}_p)^T (y - \hat{y}_p)] = E[y^T (I_n - P_p) y] = \\ &= \sigma^2 \text{tr}[I_n - P_p] + y_{cp}^T (I_n - P_p) y_{cp} = \sigma^2 (n - p) + SSB_p. \end{aligned}$$

Поэтому

$$E[C_p] \approx (n - p) + \frac{SSB_p}{\sigma^2} + 2p - n = p + \frac{SSB_p}{\sigma^2} = \Delta_p,$$

т.е. C_p является приблизительно несмещённой оценкой для Δ_p .

Задача

1. Модель $\eta_1 = \theta_1 + \theta_2 x$
2. Гипотеза принимается

Билет № 11

Вопрос 1

Одномерный анализ:

Сравнение математических ожиданий;
Сравнение дисперсий;
Оценивание параметров распределений;
Установление закона распределения;
Отбраковка данных.

Многомерный анализ:

Исследование зависимостей между признаками;
Классификация объектов;
Снижение размерности пространства признаков.

Вопрос 2

Функция $\eta(x, \theta)$ линейна по параметрам, если

$$E(y/x) = \eta(x, \theta) = \theta^T f(x),$$

Вопрос 3

Несмещенность, состоятельность, эффективность.

Вопрос 4

Использование симметричных форм критериев всегда целесообразно: во-первых, снижается чувствительность критерия к способу разделения данных на части, и, во-вторых, усредняется влияние помех, действующих на различных частях выборки.

Вопрос 5

Максимальная парная сопряженность, максимальная сопряженность

Вопрос 6

При предварительном анализе характера исследуемых зависимостей ограничиваются в основном изучением “геометрии” парных корреляционных полей. Под корреляционным полем переменных (u, v) понимается графическое представление имеющихся измерений $(u_1, v_1), (u_2, v_2), \dots, (u_n, v_n)$ этих переменных в плоскости (u, v) .

В результате такого анализа обычно получают формулировку нескольких рабочих гипотез об общем виде искомой зависимости.

Вопрос 7

Частный случай 2. Проверка незначимости уравнения регрессии.

Пусть задана линейная модель наблюдения:

$$y_i = \Theta_0 + \Theta_1 x_{1,i} + \dots + \Theta_{m-1} x_{m-1,i} + \varepsilon_i, \quad i = \overline{1, n}$$

и требуется установить является ли регрессия с заданными регрессорами незначимой, т.е. мы хотим проверить гипотезу $H: \Theta_1 = \Theta_2 = \dots = \Theta_{m-1} = 0$ или другими словами сравнить данную модель с моделью среднего $y = \Theta_0$.

Гипотеза H имеет вид: $A\Theta = 0$, где $A = [0 | I_{m-1}]$ - $(m-1) \times m$ матрица ранга $m-1$.

Применима общая теория с $q = m-1$, $RSS = y^T y - \hat{\Theta}^T X^T y$ и

$$RSS_H = \min_{\Theta_0} \sum_{i=1}^n (y_i - \Theta_0)^2 = \sum_{i=1}^n (y_i - \bar{y})^2 = y^T y - n\bar{y}^2.$$

Вопрос 8

Data Mining (или Knowledge Discovery In Data – обнаружение знаний в данных) – исследование и обнаружение машиной в сырых данных скрытых знаний, которые ранее были не известны, нетривиальны, практически полезны, доступны для интерпретации человеком.

Данные: цифровые данные, текст, изображение, видео, звук.

Методы Data Mining находятся на стыке различных направлений информационных технологий: прикладной статистики, баз данных, нейронных сетей, нечетких множеств, генетических алгоритмов, искусственного интеллекта и др.

Основные задачи **Data Mining**: классификация, построение зависимостей, кластеризация, ассоциативные правила, визуализация, поиск исключений.

Вопрос 9

Для линейной модели с двумя факторами при 4 наблюдениях в углах квадрата информационная матрица будет диагональной с диагональю 4. При 8 наблюдениях (повторы) диагональ будет равна 8 и т.д.

Вопрос 10

Условие минимума известно – оно определяется как

$$X^T \underbrace{X\hat{\theta}}_{\hat{y}} = X^T y.$$

Отсюда $X^T(y - X\hat{\theta}) = 0$ и далее $X^T(y - \hat{y}) = 0$, т.е. вектор остатков $\hat{e} = y - \hat{y}$ ортогонален Ω .

Вопрос 11

Поскольку ε_{t-1} и v_t независимы, то

$$E(\varepsilon_t^2) = E(\rho\varepsilon_{t-1} + v_t)^2 = \rho^2 E(\varepsilon_{t-1}^2) + E v_t^2 = \rho^2 E(\varepsilon_{t-1}^2) + \sigma_v^2.$$

Будем предполагать, что ошибки наблюдения ε_t имеют одинаковую дисперсию, т.е. .

$E(\varepsilon_t^2) = E(\varepsilon_{t-1}^2)$. С учетом этого можно получить

$$\sigma_\varepsilon^2 = E(\varepsilon^2) = E(\varepsilon_t^2) = E(\varepsilon_0^2) = \sigma_v^2 / (1 - \rho^2).$$

Относительно ковариации можно получить, умножая (3.6) на ε_{t-1} :

$$E(\varepsilon_t \varepsilon_{t-1}) \equiv \text{cov}(\varepsilon_t \varepsilon_{t-1}) = \rho \sigma_\varepsilon^2$$

и далее $\text{cov}(\varepsilon_t \varepsilon_{t-2}) = \rho^2 \sigma_\varepsilon^2, \dots, \text{cov}(\varepsilon_t \varepsilon_{t-m}) = \rho^m \sigma_\varepsilon^2$.

Таким образом, можно окончательно выписать ковариационную матрицу для ε :

$$V = \frac{\sigma_v^2}{1 - \rho^2} \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ . & . & . & \dots & . \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{pmatrix} = \sigma_v^2 \Omega.$$

Вопрос 12

Переменные, которым соответствуют элементы вектора y для минимального собственного числа с значимо отличающимися от 0 значениями и порождают мультиколлинеарность.

Вопрос 13

Монотонность коэффициента детерминации от числа включенных в модель регрессоров легко доказывается, если обратиться к его выражению:

$$R^2 = 1 - \frac{RSS}{\sum (y_i - \bar{y})^2}$$

Вопрос 14

При наличии повторной выборки можно предложить критерий селекции, в основу которого положена идея идеального внешнего критерия. При вычислении идеального внешнего критерия вместо зашумленных значений отклика используются точные (незашумленные) значения. Однако в реальной ситуации эти значения нам не известны. В случае повторной выборки вместо неизвестных значений вектора можно использовать и их оценку в виде выборочного среднего $\bar{y}$:

$$\bar{y}^T = (\bar{y}_A^T, \bar{y}_B^T), \bar{y}_{A_i} = (y_{A_i} + y_{B_i})/2, i = 1, \dots, n_A, \bar{y}_B = \bar{y}_A.$$

С использованием вектора $\bar{y}$ можно предложить следующий субидеальный критерий стабильности:

$$\tilde{S}^2 = \left\| \bar{y}_A - X_A \hat{\theta}_A \right\|^2 + \left\| \bar{y}_B - X_B \hat{\theta}_B \right\|^2.$$

Вопрос 15

Пусть $\tilde{\theta} = C^T y$ - произвольная несмещенная оценка, т.е.

$$E\tilde{\theta} = E(C^T y) = EC^T (X\theta + \varepsilon) = C^T X\theta = \theta.$$

Таким образом для C должно выполняться

$$C^T X = I. \quad (1.6)$$

Обозначим $H = C^T - (X^T X)^{-1} X^T$. Тогда из (1.6)

$$HX = (C^T - (X^T X)^{-1} X^T)X = C^T X - (X^T X)^{-1} X^T X = C^T X - I = 0,$$

т.е.

$$HX = 0.$$

Определим смещение оценки $\tilde{\theta}$:

$$\tilde{\theta} - \theta = C^T y - \theta = C^T X\theta + C^T \varepsilon - \theta = C^T \varepsilon.$$

Вычислим матрицу ковариаций для оценки $\tilde{\theta}$:

$$\begin{aligned} \text{Cov}(\tilde{\theta}) &\equiv D(\tilde{\theta}) = E(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^T = E(C^T \varepsilon \varepsilon^T C) = \sigma^2 C^T C = \\ &= \sigma^2 [H + (X^T X)^{-1} X^T][H^T + X(X^T X)^{-1}] = \\ &= \sigma^2 [HH^T + HX(X^T X)^{-1} + (X^T X)^{-1} X^T H^T + (X^T X)^{-1} X^T X(X^T X)^{-1}] = \\ &= \sigma^2 HH^T + \sigma^2 (X^T X)^{-1} = \sigma^2 HH^T + D(\hat{\theta}) \end{aligned}$$

Действительно второе слагаемое в полученной сумме есть дисперсия оценки $\hat{\theta}$, поскольку: $\hat{\theta} - \theta = (X^T X)^{-1} X^T \varepsilon$ и

$$D(\hat{\theta}) = E(\hat{\theta} - E\hat{\theta})(\hat{\theta} - E\hat{\theta})^T = E[(X^T X)^{-1} X^T \varepsilon \varepsilon^T X (X^T X)^{-1}] = \sigma^2 (X^T X)^{-1} \blacksquare$$

В силу своего представления матрица HH^T - п.п.о., поэтому $D(\tilde{\theta}) \geq D(\hat{\theta}) \blacksquare$

Для того чтобы $HH^T = 0$ необходимо чтобы $H = 0$. Отсюда получаем, что $C^T = (X^T X)^{-1} X^T$ и $\tilde{\theta} = (X^T X)^{-1} X^T y = \hat{\theta} \blacksquare$

Задача

1. Модель $\eta_1 = \theta_1 + \theta_2 x$
2. Гипотеза отвергается

Билет №12

Вопрос 1

Методы статистического исследования зависимостей составляют содержание отдельных частей многомерного статистического анализа.

Они формировались и развивались с учетом специфики анализируемых моделей, обусловленной природой изучаемых переменных.

Соответствующая спецификация этих разделов отражена в следующей таблице.

№	Природа откликов	Природа факторов	Название раздела многомерного статистического анализа
1.	Количественная	Количественная	Регрессионный анализ
2.	Количественная	Одна переменная – время	Анализ временных рядов
3.	Количественная	Качественная	Дисперсионный анализ
4.	Количественная	Смешанная	Ковариационный анализ
5.	Неколичественная	Неколичественная	Анализ ранговых корреляций
6.	Неколичественная или отсутствие	Количественная	Дисперсионный анализ, кластерный анализ
7.	Смешанная	Смешанная	Аппарат логических решающих функций

Вопрос 2

Термины «регрессионная модель», «регрессионный анализ» – устоявшиеся, хотя и не совсем удачные.

Термин «регрессия» ввел антрополог Гальтон (1822-1911) в 1885г. Он установил «закон регрессии». Имеется в виду, что «возврат» (регрессия) потомков в отношении их индивидуальных признаков к родителям происходит в «среднем». В частности он рассматривал размер семян и рост детей и родителей. У рослых родителей не обязательно еще более рослые дети и наоборот, у малорослых родителей – еще более малорослые. Он сравнивал средний рост детей Y со средним ростом их родителей X . Гальтон заметил, что рост детей у высоких (низких) родителей обычно также выше (ниже) среднего роста популяции $\mu \cong \bar{X} \cong \bar{Y}$, но при этом отклонение от μ у детей меньше, чем у родителей. Другими словами, экстремумы в следующем поколении сглаживаются, происходит возвращение назад (*регрессия*) к середине. Гальтон показал, что зависимость Y от X хорошо выражается уравнением $Y - \bar{Y} = (2/3)(X - \bar{X})$.

Вопрос 3

A1: $E(\varepsilon) = 0$; A2: $E(\varepsilon\varepsilon^T) = \sigma^2 I_n$; A3: $rg(X) = m$.

Вопрос 4

Термин "сферичность" возмущений происходит из того наблюдения, что в многомерном пространстве распределения помехи ε при $Var(\varepsilon) = \sigma^2 I_n$ (случай равнооточных наблюдений) точки с одинаковой вероятностью образуют сферы. В рассматриваемом случае неравнооточных наблюдений вместо сфер будут образовываться многомерные эллипсоиды.

Вопрос 5

Нет.

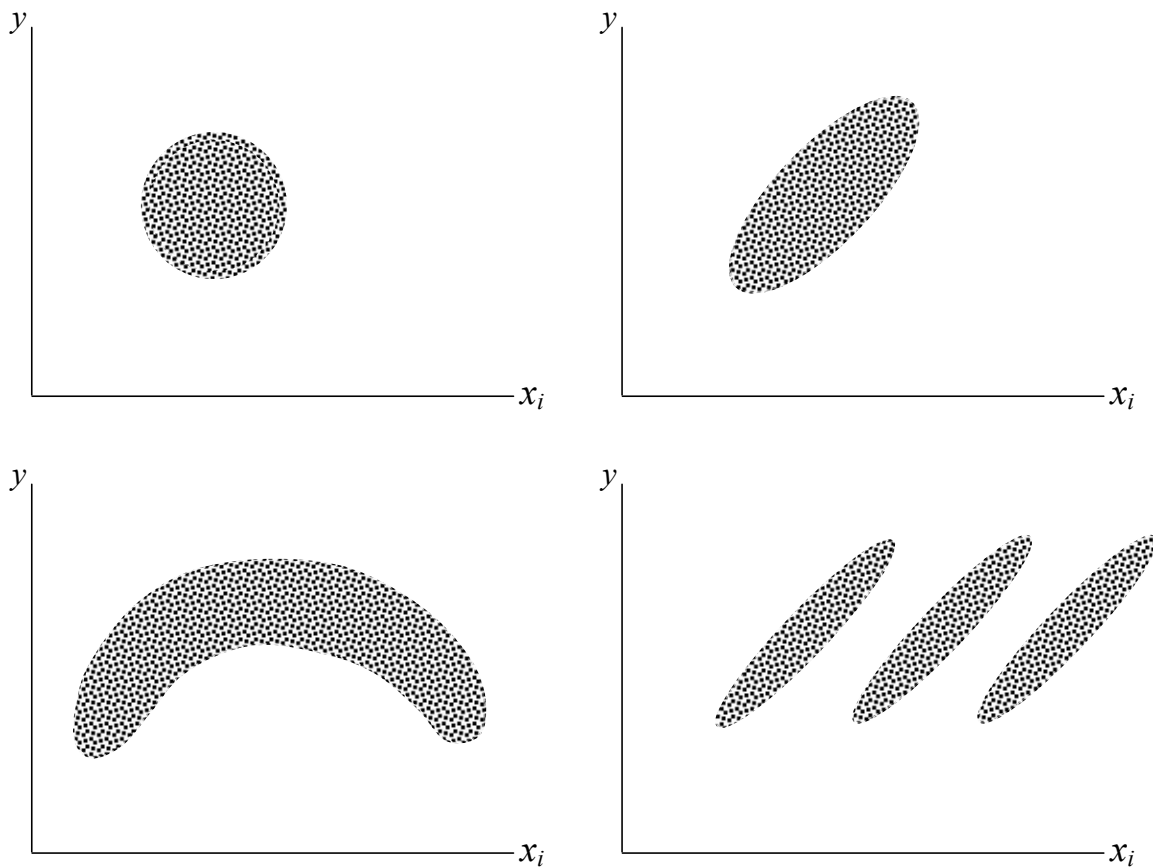
Вопрос 6

При предварительном анализе характера исследуемых зависимостей ограничиваются в основном изучением "геометрии" парных корреляционных полей. Под корреляционным полем переменных (u, v) понимается графическое представление имеющихся измерений $(u_1, v_1), (u_2, v_2), \dots, (u_n, v_n)$ этих переменных в плоскости (u, v) .

Анализ парных корреляционных полей состоит обычно в следующем:

а) построение на основании имеющихся исходных данных корреляционных полей для всевозможных пар переменных вида (x_j, x_l) и (x_j, y) — всего таких пар будет $k(k+1)/2$. Этот процесс легко автоматизируется с помощью средств современных ЭВМ;

б) визуальное прослеживание характера вытянутости каждого корреляционного поля (эллипсоидально-линейное, нелинейное-монотонное, с наличием экстремумов и т. п.);



Это примеры корреляционных полей для пассивного эксперимента.

Вопрос 7

Критерий стабильности, регулярности.

Вопрос 8

Этапы: постановочный. Информационный, корреляционный анализ, определение класса допустимых решений, анализ мультиколлинеарности, оценка параметров, проверка адекватности

Вопрос 9

Имеем

$$\hat{\theta} = (X^T X)^{-1} X^T (X\theta + \varepsilon) = \theta + (X^T X)^{-1} X^T \varepsilon;$$

$$E(\hat{\theta}) = E(\theta) + (X^T X)^{-1} X^T E(\varepsilon) = \theta$$

поскольку действует предположение А1.

Вопрос 10

Поскольку $\hat{y}$ и $\hat{e}$ из ортогональных подпространств, то ясно, что $\hat{y}^T \hat{e} = 0$. Далее, поскольку $y = \hat{y} + \hat{e}$, то $y^T y = \hat{y}^T \hat{y} + \hat{e}^T \hat{e}$.

Вопрос 11

Доступный обобщенный МНК. В случае обнаружения эффекта гетероскедастичности можно провести коррекцию оценок параметров модели с помощью взвешенного МНК с весами $\hat{\sigma}_t^{-2} = h^{-1}(z_t^T \hat{\alpha})$. Может оказаться, что для некоторых t $h^{-1}(z_t^T \hat{\alpha}) < 0$. Если число таких наблюдений невелико, то их можно просто выбросить, в противном случае использовать представление $\sigma_t^2 = \exp(h(z_t^T \alpha))$.

Вопрос 12

Можно рекомендовать использовать информационную матрицу, нормированную на след:

$$X^T X / \text{tr} X^T X.$$

Вопрос 13

В 2 раза.

Вопрос 14

Предположим, что ошибки наблюдения распределены по нормальному закону $N(0, \sigma^2 I_N)$. Пусть априорные данные о параметрах модели определены информацией из выборки A и состоят в выполнении $X_A^T X_A \theta = X_A^T y_A$. Поступают новые данные в виде выборки B . По ней производится оценивание параметров и осуществляется проверка гипотезы $H: X_A^T X_A \theta = X_A^T y_A$.

Обозначим

$$RSS = \|y_B - \hat{y}_B(B)\|^2 = (y_B - X_B \hat{\theta}_B)^T (y_B - X_B \hat{\theta}_B) \equiv \varepsilon^2(B), \quad)$$

$$RSS_H = (y_B - X_B \hat{\theta}_H)^T (y_B - X_B \hat{\theta}_H), \quad)$$

где $\hat{\theta}_H$ - оценки, полученные при ограничениях $X_A^T X_A \theta = X_A^T y_A$.

В нашем случае при условии полного ранга матрицы $X_A^T X_A$ имеем

$$\hat{\theta}_H = \hat{\theta}_A = (X_A^T X_A)^{-1} X_A^T y_A.$$

Статистика F для проверки гипотезы H будет следующая

$$F = \frac{(RSS_H - RSS) / p}{RSS / (N_B - p)},$$

где p – число параметров в модели. С учетом ранее введенных обозначений F -статистику можно переписать

$$F = \frac{(\Delta_{reg}^2(B) - \varepsilon^2(B)) / p}{\varepsilon^2(B) / (N_B - p)} = \frac{\Delta_{reg}^2(B)}{\varepsilon^2(B)} - \frac{N_B - p}{p}.$$

Вопрос 15

Под состоятельностью понимается сходимость по вероятности

$$\lim_{N \rightarrow \infty} P \left[\left(\hat{\theta}_N - \theta \right)^T \left(\hat{\theta}_N - \theta \right) \geq \gamma \right] = 0.$$

Состоятельность оценки (точнее – последовательности оценок) означает концентрацию вероятностной массы около истинного значения параметра с ростом объема выборки. В многомерном случае требуется, чтобы состоятельность выполнялась для каждого компонента θ_i вектора θ .

Будем рассматривать состоятельность в среднем квадратичном:

$$E(\hat{\theta}_N - \theta)^T (\hat{\theta}_N - \theta) \rightarrow 0, N \rightarrow \infty.$$

Сходимость в среднеквадратичном влечет сходимость по вероятности. Это сразу вытекает из неравенства Чебышева

$$P[|\hat{\theta}_N - \theta| \geq \gamma] \leq \frac{1}{\gamma} E(\hat{\theta}_N - \theta)^2.$$

Величина $E(\hat{\theta}_N - \theta)^T (\hat{\theta}_N - \theta)$ есть след матрицы ковариаций $E(\hat{\theta}_N - \theta)(\hat{\theta}_N - \theta)^T \equiv D(\hat{\theta}_N)$.

Таким образом для выполнения условия

$tr D(\hat{\theta}_N) \equiv E(\hat{\theta}_N - \theta)^T (\hat{\theta}_N - \theta) \rightarrow 0, N \rightarrow \infty$ необходимо потребовать выполнение условия $D(\hat{\theta}_N) \rightarrow 0, N \rightarrow \infty$.

Поскольку $D(\hat{\theta}_N) = \sigma^2 (\sum_{i=1}^N f(x_i) f^T(x_i))^{-1} = \sigma^2 (X^T X)^{-1}$ необходимо чтобы матрица $X^T X$

бесконечно возрастала с ростом N . Поскольку $f(x_i) f^T(x_i)$ - п.п.о., то всегда найдется такое

k , что $\sigma^2 \sum_{i=1}^N f(x_i) f^T(x_i) < \sigma^2 \sum_{i=1}^{N+k} f(x_i) f^T(x_i)$,

что означает неограниченный рост $X^T X$ ■

Задача

1. Модель $\eta_1 = \theta_1 + \theta_2 x^2$
2. Гипотеза принимается

Билет №13

Вопрос 1

При активном эксперименте исследователь сам устанавливает необходимые комбинации значений входных факторов.

Вопрос 2

Адекватность модели – способность описывать поведение объекта (системы) в среднем в условиях того, что экспериментальные наблюдения за откликом подвержены влиянию случайных помех.

Вопрос 3

Доверительное оценивание для отдельного параметра

Доверительный интервал для отдельного параметра также можно получить исходя из общего выражения:

$$(\Theta_j - \hat{\Theta}_j)^T \frac{1}{d_{jj}} (\Theta_j - \hat{\Theta}_j) \leq \hat{\sigma}^2 F_{\alpha, 1, n-m}$$

Учитывая, что $((X^T X)^{-1})_{jj} \sigma^2 = \sigma^2 (\hat{\Theta}_j)$ - дисперсия оценки отдельного параметра, и что ,

где $t_{\alpha/2, f_R}$ - t -критерий Стьюдента, $f_R = n-m$, то можно записать в виде: $\frac{|\Theta_j - \hat{\Theta}_j|}{\sigma(\hat{\Theta}_j)} \leq t_{\alpha/2, f_R}$.

Или в виде двухстороннего неравенства:

$$\hat{\Theta}_j - t_{\alpha/2, f_R} \sigma(\hat{\Theta}_j) \leq \Theta_j \leq \hat{\Theta}_j + t_{\alpha/2, f_R} \sigma(\hat{\Theta}_j).$$

Вопрос 4

Выделяющиеся точки графика: некоторые из остатков могут по абсолютной величине сильно превосходить все остальные остатки. Если максимальное значение остатка $|d_{max}| > \delta$, где δ - некоторое заранее выбранное число, то наблюдение, имеющее такой остаток, является аномальным. Величина остатка аномального наблюдения зависит от объема выборки n : при $n > 20$ его можно обнаружить по правилу «трех сигм» ($\delta = 3$), при $n < 20$, например, по критерию Смирнова - Граббса. В этом случае вычисляемая статистика имеет вид

$$D = (d_{\max} - \bar{d}) / S_d, \text{ где } \bar{d} - \text{среднее значение остатков, } S_d^2 = 1/n \left(\sum_{i=1}^n (d_i - \bar{d})^2 \right).$$

Вопрос 5

Ридж-оценка имеет минимальную длину в классе оценок с заданным значением суммы квадратов отклонений (RSS).

Вопрос 6

Статистика Мэллоуса, коэффициент детерминации, критерий Акаике.

Вопрос 7

Для критерия Δ_{abs}^2 , например, его среднее значение будет определяться величиной

$$\bar{\Delta}_{abs}^2 = u^T D_{\Delta_{abs}(p)} u + \sigma^2 \cdot p.$$

С увеличением σ^2 минимум критерия в его зависимости от сложности модели смещается влево, т.е. выбираются все более простые модели. Абсолютно помехоустойчивый критерий получил свое название по причине того, что второе слагаемое для него менее всего возрастает при увеличении дисперсии помехи. Другими словами, при увеличении дисперсии помехи минимум данного критерия с наименьшей скоростью смещается влево в сторону все более простых моделей.

Вопрос 8

На первом шаге выбирается некоторое число лучших моделей, состоящих из одного регрессора. На втором этапе к этим моделям добавляются различные регрессоры по одному, и отбираются модели с двумя аргументами, улучшающие величину критерия, и т.д. до его ухудшения или до построения полной модели.

На последнем шаге можно из оставшихся лучших моделей выбрать одну по другому критерию.

Вопрос 9

Для линейной модели с двумя факторами при 4 наблюдениях в углах квадрата информационная матрица будет диагональной с диагональю 4. При 8 наблюдениях (повторы) диагональ будет равна 8 и т.д.

Вопрос 10

Имеем $X^T(y - X\hat{\theta}) = 0$ и далее $X^T(y - \hat{y}) = 0$, т.е. вектор остатков $\hat{e} = y - \hat{y}$ ортогонален Ω . В тоже время справедливо представление $y = \hat{y} + \hat{e}$. Становится понятным, что $\hat{y}$ есть проекция y на $R(X)$, а именно $\hat{y} = Py$. Можно рассматривать также подпространство $R^\perp(X)$ – ортогональное дополнение для $R(X)$ в $R(X)$. Поскольку

$$\hat{e} = y - \hat{y} = y - X\hat{\theta} = (I_n - X(X^T X)^{-1} X^T)y = (I_n - P)y,$$

то $M = I_n - P$ – есть матрица проектирования вектора y на подпространство $R^\perp(X)$.

Вопрос 11

$$\hat{\theta} = (X^T X)^{-1} X^T y. \quad \hat{\theta} = (n)^{-1} X^T y = \sum_{i=1}^n y_i / n$$

Вопрос 12

Для наглядности рассмотрим случай $k=2$.

$$y = \theta_1 x_1 + \theta_2 x_2 + \varepsilon.$$

Пусть для удобства мы пронормировали вектора

$$\underline{x}_1^T = (x_{11}, \dots, x_{1n}), \quad \underline{x}_2^T = (x_{21}, \dots, x_{2n}): \quad \underline{x}_1' = \frac{\underline{x}_1}{\|\underline{x}_1\|}, \quad \underline{x}_2' = \frac{\underline{x}_2}{\|\underline{x}_2\|}.$$

Тогда

$$X^T X = \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}, \text{ где } r_{12} = \cos(\underline{x}_1, \underline{x}_2) = \sum \frac{x_{1t} x_{2t}}{\sqrt{\sum x_{1t}^2 \cdot \sum x_{2t}^2}}.$$

Тогда

$$(X^T X)^{-1} = \begin{pmatrix} \frac{1}{(1-r_{12}^2)} & -\frac{r_{12}}{(1-r_{12}^2)} \\ -\frac{r_{12}}{(1-r_{12}^2)} & \frac{1}{(1-r_{12}^2)} \end{pmatrix} \text{ и } \sigma^2(\hat{\theta}_1) = \sigma^2(\hat{\theta}_2) = \frac{\sigma^2}{(1-r_{12}^2)}.$$

Если между переменными x_1 и x_2 существует некая линейная зависимость, т.е. r_{12}^2 близко к 1, то, как следует из последнего выражения, дисперсии оценок параметров будут иметь большие значения. При $r_{12} \rightarrow 1 \Rightarrow \sigma^2(\hat{\theta}_1) \rightarrow \infty$.

Вопрос 13

2) $\hat{\beta}(\lambda)$ - является линейным преобразованием МНК-оценки и является смещённой;

Действительно,

$$Z^T y = (I + \lambda(Z^T Z)^{-1})^{-1} (Z^T Z)^{-1} Z^T y = B\hat{\beta},$$

где $B = (I + \lambda(Z^T Z)^{-1})^{-1}$ - матрица линейного преобразования и $B \neq I$, что приводит к смещению оценок $\hat{\beta}(\lambda)$ в отличие от $\hat{\beta}$, которые несмещены.

Вопрос 14

При переусложнении модели имеется дополнительная возможность описывать не только среднее поведение системы, но и флуктуации, обусловленные влиянием помех. На новых тестовых данных помеха уже будет другая, что и создаст условия увеличения ошибки предсказания.

Вопрос 15

Введем в рассмотрение вектор остатков: -

$$\hat{e} = y - \hat{y} = y - X\hat{\theta} = (I_n - X(X^T X)^{-1} X^T) y = (I_n - P) y,$$

где P - матрица проектирования: $\hat{y} = Py$, т.е. y проектируется на Ω ; $(I_n - P)$ - матрица ортогонального проектирования на $\Omega^\perp$, где $\Omega = R[X]$ - пространство образов, порожаемое столбцами матрицы X .

Матрицы P и $I_n - P$ симметричны и идемпотентны, т.е. $P^2 = P$, $(I_n - P)^2 = (I_n - P)$.

Поскольку матрица $I_n - P$ симметрична и идемпотентна, то $\text{rang}[I_n - P] = \text{tr}[I_n - P] = n - \text{tr}[P]$ т.е. у идемпотентной матрицы ранга r имеется r собственных чисел равных 1, а остальные $n - r$ равны 0. В нашем случае:

$$\text{tr}[P] = \text{tr}[X(X^T X)^{-1} X^T] = \text{tr}[X^T X (X^T X)^{-1}] = \text{tr}[I_m] = m.$$

Введем в рассмотрение статистику

$$\hat{e}^T \hat{e} = (y - X\hat{\theta})^T (y - X\hat{\theta}) = y^T (I_n - P)^T (I_n - P) y = y^T (I_n - P)^2 y.$$

Теорема о математическом ожидании квадратичной формы от случайной величины:

Пусть Z - случайный вектор размера $n \times 1$, A - симметричная $n \times n$ матрица. Если $E(Z) = \xi$ и $D(Z) = \Sigma$, то $E[Z^T A Z] = \text{tr}[A \Sigma] + \xi^T A \xi$.

Применительно к нашему случаю имеем:

$$E[\hat{e}^T \hat{e}] = E[y^T (I_n - P) y] = \sigma^2 \text{tr}(I_n - P) + \theta_{\text{ндо}}^T X^T (I_n - P) X \theta_{\text{ндо}} = \sigma^2 (n - m).$$

Примечание. Второе слагаемое здесь равно нулю поскольку $(I_n - P)X = 0$.

Отсюда заключаем, что статистика $\hat{\sigma}^2 = \frac{\hat{e}^T \hat{e}}{n - m}$ - несмещенная оценка для σ^2 .

Задача

1. Модель $\eta_1 = \theta_1 + \theta_2 x^2$
2. Гипотеза отвергается

Билет №14

Вопрос 1

НЛО, МНК, ММП, Ридж-оценивание, ОМНК

Вопрос 2

Для возможности обращения информационной матрицы

Вопрос 3

Для ее вычисления необходимо обратить матрицу коэффициентов корреляций, которая может быть плохо обусловлена в условиях мультиколлинеарности.

Вопрос 4

Ридж-оценки впервые были введены А. Гоэрлом при исследовании регрессии 2-го порядка (ridge – гребень, хребет). Название "ридж-оценки" навеяно следующими наблюдениями. При сильной мультиколлинеарности линии равного уровня доверительного эллипсоида для разного уровня значимости (доверительной вероятности) становятся сильно вытянутыми и визуально похожими на хребет на географической карте. Данные оценки использовались именно в условиях мультиколлинеарности и поэтому получили название ридж-оценок.

Вопрос 5

Ошибка спецификации модели – это недобор или перебор регрессоров.

Вопрос 6

Критерии точности.

Типичным критерием этого вида является критерий регулярности:

$$\begin{aligned}\Delta_{reg}^2(B) &= \Delta^2(B/A) = \|y_B - \hat{y}_B(A)\|^2 = \\ &= (y_B - X_B \hat{\theta}_A)^T (y_B - X_B \hat{\theta}_A) = \|y_B - X_B \hat{\theta}_A\|^2\end{aligned}$$

где запись $\Delta^2(B/A)$ означает «ошибка на выборке B модели, коэффициенты которой получены на A », $\hat{\theta}_A = (X_A^T X_A)^{-1} X_A^T y_A$, $\hat{y}_B(A) = X_B \hat{\theta}_A$. Этот критерий несимметричен.

Вопрос 7

Случай недобора.

Пусть истинное уравнение (4.2), а мы оцениваем регрессию

$$y_u = \gamma_1 \omega_{u1} + \dots + \gamma_k \omega_{uk} + \xi_u \quad (4.5)$$

или

$$y = W\gamma + \xi, \quad X = \begin{bmatrix} W & V \end{bmatrix}$$

Другими словами, $\omega_1 = x_1, \dots, \omega_k = x_k$ входят в регрессию (4.5), а остальные $m - k$ переменных $\omega_{k+1}, \dots, \omega_{m-k} = x_m$ в регрессии отсутствуют.

Докажем, что тогда используемая оценка

$$\hat{\gamma} = (W^T W)^{-1} W^T y \quad (4.6)$$

в общем случае является смещённой в том смысле, что $E\hat{\gamma} \neq \alpha_1$.

Действительно,

$$E\hat{\gamma} = (W^T W)^{-1} W^T E y = (W^T W)^{-1} W^T X \alpha = (W^T W)^{-1} W^T \begin{bmatrix} W & V \end{bmatrix} \cdot \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix},$$

где α_1 и α_2 - подвекторы вектора α размерности k и $m - k$ соответственно.

Перемножая, получим

$$E\hat{\gamma} = \begin{bmatrix} I_k & (W^T W)^{-1} W^T V \end{bmatrix} \cdot \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \alpha_1 + (W^T W)^{-1} W^T V \alpha_2 \quad \blacksquare$$

Оценка $\hat{y}$ будет несмещённой, когда матрицы W и V ортогональны, то есть $W^T V = 0$.

Можно показать также, что МНК оценка в случае недобора является несостоятельной.

Вопрос 8

Для доказательства несмещенности оценок.

Вопрос 9

По причинам неадекватности модели, обнаружения мультиколлинеарности.

Вопрос 10

$$P = X(X^T X)^{-1} X^T. \quad P^2 = X(X^T X)^{-1} X^T X(X^T X(X^T X)^{-1} X^T)^{-1} X^T = P$$

Вопрос 11

Точечную оценку дисперсии ошибки эксперимента в случае, когда имеется повторные опыты лучше всего определять через сумму (2.23) по формуле

$\hat{\sigma}_E^2 = SS_E / f_E$, где f_E число степеней свободы суммы SS_E , равное

$f_E = N - N_c = \sum_{v=1}^{N_c} (r_v - 1)$. Оценка $\hat{\sigma}_E^2$ есть не что иное, как объединенная оценка

выборочных дисперсий ошибки опыта, которые можно найти по результатам отдельных серий повторных опытов. Оценка дисперсии ошибки в v -ой серии повторных опытов равна

$$\hat{\sigma}_{Ev}^2 = \frac{SS_{Ev}}{f_{Ev}}, \text{ где } SS_{Ev} = \sum_{j=1}^{r_v} (y_{vj} - \bar{y}_v)^2, \quad f_{Ev} = r_v - 1.$$

Отсюда следует, что объединенная оценка по всем сериям имеет вид

$$\hat{\sigma}_E^2 = \sum_{v=1}^{N_{\bar{a}}} \frac{f_{Ev}}{f_E} \hat{\sigma}_{Ev}^2 = \sum_{v=1}^{N_{\bar{a}}} SS_{Ev} / f_E = SS_E / f_E.$$

Вопрос 12

. Рассмотрим случай когда в модели присутствует свободный член. В этом случае сумма остатков равно нулю, т.е.

$$\sum (y_i - \hat{y}_i) = 0. \quad (a)$$

Покажем это. Для этого сумму квадратов отклонений для модели со свободным членом

$$\sum (y_i - \theta_0 - \theta_1 x_{1i} - \dots - \theta_{m-1} x_{m-1,i})^2$$

продифференцируем по θ_0 :

$$\sum (y_i - \hat{\theta}_0 - \hat{\theta}_1 x_{1i} - \dots - \hat{\theta}_{m-1} x_{m-1,i}) = 0 \text{ или } \sum (y_i - \hat{y}_i) = 0.$$

Вопрос 13

Ясно, что в реальных ситуациях редко бывает так, что параметр автокорреляции ρ известен. При неизвестном ρ процедуры оценивания параметров θ носят итеративный характер.

Процедура Кохрейна-Оркатта.

1. Для исходной модели оцениваются $\hat{\theta}$ и $\hat{e}$;

2. Рассматривается регрессия $\hat{e}_t = \rho \hat{e}_{t-1} + v_t$ и вычисляется оценка $\hat{\rho}$;
3. Проводится ОМНК оценивание исходной модели с $\hat{\Omega}^{-1}$, где в качестве ρ используется его оценка $\hat{\rho}$;
4. Вычисляется новый вектор остатков $\hat{e}$;
5. Процедура повторяется с п.2.

Останов производится после нескольких итераций или после того как очередное $\hat{\rho}$ мало отличается от предыдущего.

Процедура Хилдрета-Лу.

По схеме ОМНК производится решение следующей экстремальной задачи:

$$\hat{\rho} = \text{Arg} \min_{\rho \in [-1, +1]} \text{RSS}.$$

Попутно при этом мы получаем оценки для вектора параметров θ .

Вопрос 14

$$p_{ii} = f^T(x_i)(X^T X)^{-1} f(x_i)$$

. На практике руководствуются следующими правилами. Значения $p_{ii} \leq 0,2$ позволяют считать надежными данные наблюдения, значения $0,2 \leq p_{ii} \leq 0,5$ относят данное наблюдение к классу рискованных, а при значениях $p_{ii} > 0,5$ i -е наблюдение можно принять за точку разбалансировки.

Вопрос 15

Пусть как и ранее модель наблюдения имеет вид:

$$y = X\Theta + \varepsilon,$$

где X - матрица $n \times m$ ранга m .

Предположим, что мы хотим найти оценки параметра Θ , минимизируя $\varepsilon^T \varepsilon$, при наличии линейных ограничений $A\Theta = c$, где A - известная $q \times m$ матрица ранга $q < m$, а c - известный вектор $q \times 1$.

Применим метод множителей Лагранжа. Рассмотрим целевую функцию:

$$r = \varepsilon^T \varepsilon + (\Theta^T A^T - c^T) \lambda,$$

где $\lambda^T = (\lambda_1, \dots, \lambda_q)$ - вектор множителей Лагранжа. Решим уравнения:

$$A\Theta = c \tag{2.1}$$

$$\frac{\partial r}{\partial \Theta} = -2X^T y + 2X^T X \Theta + A^T \lambda = 0. \tag{2.2}$$

Решения уравнений (2.1) и (2.2) обозначим через $\hat{\lambda}_H$ и $\hat{\Theta}_H$. Тогда из (2.2) получаем:

$$\hat{\Theta}_H = (X^T X)^{-1} X^T y - \frac{1}{2} (X^T X)^{-1} A^T \hat{\lambda}_H = \hat{\Theta} - \frac{1}{2} (X^T X)^{-1} A^T \lambda_H, \tag{2.3}$$

а из (2.1):

$$c = A\hat{\Theta}_H = A\hat{\Theta} - \frac{1}{2} A(X^T X)^{-1} A^T \hat{\lambda}_H.$$

Поскольку матрица $(X^T X)^{-1}$ - п.п.о., то матрица $A(X^T X)^{-1} A^T$ будет п.п.о. ранга q .
Следовательно

$$-\frac{1}{2} \hat{\lambda}_H = [A(X^T X)^{-1} A^T]^{-1} (C - A\hat{\Theta}).$$

Окончательно подставляя в (2.3):

$$\hat{\Theta}_H = \hat{\Theta} + (X^T X)^{-1} A^T [A(X^T X)^{-1} A^T]^{-1} (C - A\hat{\Theta}). \quad (2.4)$$

Задача

1. Модель $\eta_1 = \theta_1 x + \theta_2 x^2$
2. Гипотеза отвергается

Билет №15

Вопрос 1

Структура модели – это набор регрессоров.

Вопрос 2

Оценки параметров распределены по нормальному закону с мат.ожиданием равным истинным значениям параметров и дисперсией $\sigma^2 (X^T X)^{-1}$

Вопрос 3

Это проверка того, что модель системы не изменилась со временем.

Вопрос 4

- 2) $\hat{\beta}(\lambda)$ - является линейным преобразованием МНК-оценки и является смещённой;
Действительно,

$$Z^T y = (I + \lambda(Z^T Z)^{-1})^{-1} (Z^T Z)^{-1} Z^T y = B \hat{\beta},$$

где $B = (I + \lambda(Z^T Z)^{-1})^{-1}$ - матрица линейного преобразования и $B \neq I$, что приводит к смещению оценок $\hat{\beta}(\lambda)$ в отличие от $\hat{\beta}$, которые несмещены.

Вопрос 5

Под помехоустойчивостью внешнего критерия будем понимать его способность выбирать из всего множества моделей, построенных по зашумленным данным, такую модель оптимальной сложности, которая достаточно точно восстанавливает не зашумленный выход объекта.

Вопрос 6

$$I - P = I - X(X^T X)^{-1} X^T.$$

$$(I - P)^2 = I - 2P + P^2 = \\ = I - 2X(X^T X)^{-1} X^T + X(X^T X)^{-1} X^T X(X^T X(X^T X)^{-1} X^T)^{-1} X^T = I - P$$

Вопрос 7

Частный случай 1: проверка незначимости отдельного параметра Н: $\Theta_j = c$ или Н:

$a^T \Theta = c$, где a^T - вектор-строка, в которой на j -ом месте стоит 1, на остальных местах - нули. Обозначим $d_{jj} = ((X^T X)^{-1})_{jj}$ - j -ый диагональный элемент. Тогда

$$F = \frac{(a^T \hat{\Theta} - c)^2}{\hat{\sigma}^2 d_{jj}} = \frac{(\hat{\Theta}_j - c)^2}{\hat{\sigma}^2 d_{jj}}$$

имеет при гипотезе Н распределение $F_{1, n-m}$. т.е. если $F \leq F_{\alpha, 1, n-m}$, где $F_{\alpha, 1, n-m}$ - критическая точка, то гипотеза о незначимости не отклоняется.

Вопрос 8

Доступный обобщенный МНК. В случае обнаружения эффекта гетероскедастичности можно провести коррекцию оценок параметров модели с помощью взвешенного МНК с весами $\hat{\sigma}_i^{-2} = h^{-1}(z_i^T \hat{\alpha})$. Может оказаться, что для некоторых t $h^{-1}(z_t^T \hat{\alpha}) < 0$. Если число таких наблюдений невелико, то их можно просто выбросить, в противном случае использовать представление $\sigma_i^2 = \exp(h(z_i^T \alpha))$.

Вопрос 9

В качестве критерия величины смещения (функции риска) выберем среднюю сумму квадратов ошибок (ССКО)

$$L_{\tilde{\theta}}(\theta) = E(\tilde{\theta} - \theta)^T (\tilde{\theta} - \theta),$$

где $\tilde{\theta}$ - некоторая оценка параметров $\theta \in R^m$.

Для МНК оценки

$$L_{\hat{\theta}}(\theta) = \sigma^2 \text{tr}(X^T X)^{-1} \rightarrow \infty \text{ при } \lambda_1 \rightarrow 0,$$

т.е. МНК оценка в условиях усиления коллинеарности ($\lambda_1 \rightarrow 0$) не обеспечивает ограниченность ССКО. Это будет происходить и в условиях увеличения числа наблюдений. В нормальных условиях (при отсутствии мультиколлинеарности) $\text{tr}(X^T X)^{-1} \rightarrow 0$ при $N \rightarrow \infty$, т.е. теоретически гарантируется состоятельность в среднеквадратичном.

Вопрос 10

Пусть истинное уравнение (4.2), а мы оцениваем регрессию

$$y_u = \gamma_1 \omega_{u1} + \dots + \gamma_k \omega_{uk} + \xi_u \quad (4.5)$$

или

$$y = W\gamma + \xi, \quad X = \begin{bmatrix} W & V \end{bmatrix}$$

Другими словами, $\omega_1 = x_1, \dots, \omega_k = x_k$ входят в регрессию (4.5), а остальные $m - k$ переменных $\omega_{k+1}, \dots, \omega_{m-k} = x_m$ в регрессии отсутствуют.

Докажем, что тогда используемая оценка

$$\hat{y} = (W^T W)^{-1} W^T y \quad (4.6)$$

в общем случае является смещённой в том смысле, что $E\hat{y} \neq \alpha_1$.

Действительно,

$$E\hat{y} = (W^T W)^{-1} W^T E y = (W^T W)^{-1} W^T X \alpha = (W^T W)^{-1} W^T [W \ V] \cdot \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix},$$

где α_1 и α_2 - подвекторы вектора α размерности k и $m - k$ соответственно.

Перемножая, получим

$$E\hat{y} = \begin{bmatrix} I_k & (W^T W)^{-1} W^T V \end{bmatrix} \cdot \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \alpha_1 + (W^T W)^{-1} W^T V \alpha_2 \quad \blacksquare$$

Оценка $\hat{y}$ будет несмещённой, когда матрицы W и V ортогональны, то есть $W^T V = 0$.

Можно показать также, что МНК оценка в случае недобора является несостоятельной.

Вопрос 11

В таблице исходных данных могут присутствовать наблюдения (например, наблюдение под номером i), сильно влияющие на оценки параметров. Если эти наблюдения связаны с матрицей X , то наблюдение x_i называется влияющим или точкой разбалансировки. Точки разбалансировки не обязательно дают большие остатки в соответствующих наблюдениях, эти большие остатки могут появиться в других наблюдениях.

Вопрос 12

В 4 раза.

Вопрос 13

Перебор регрессоров приводит к неэффективности оценок. Недобор – к смещению.

Вопрос 14

Критерий непротиворечивости, абсолютно помехоустойчивый критерий.

Вопрос 15

Рассмотрим более общий случай, когда среди опытов, входящих в эксперимент, имеются повторные.

Без ограничения общности можно сгруппировать опыты по сериям так чтобы каждая из них содержала все повторные опыты, которые проведены при одних и тех же условиях.

Положим, что всего имеются N_c серий и в каждую v -ю серию входят r_v - повторных опытов,

$v = \overline{1, N_c}$. Общее число опытов $N = \sum_{v=1}^{N_c} r_v$

Как и раньше рассматриваем линейную модель наблюдения

$$y(x) = f^T(x) \theta + \varepsilon \quad (2.14)$$

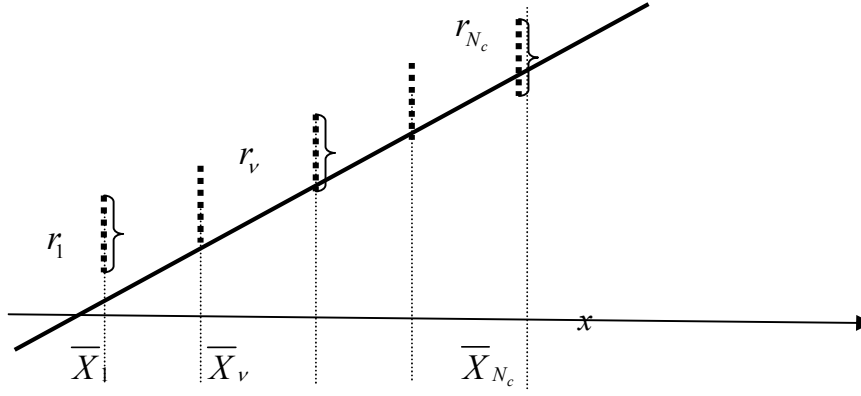
или в матричной форме

$$y = X \theta + \varepsilon. \quad (2.15)$$

Будем считать, что элементы векторов y, ε и строки матрицы X сгруппированы по сериям.

y





Введем матрицу усреднения

$$S = \begin{pmatrix} r_1^{-1} I_{r_1}^T & & & 0 \\ & r_2^{-1} I_{r_2}^T & & \\ & & \ddots & \\ 0 & & & r_{N_c}^{-1} I_{r_{N_c}}^T \end{pmatrix} \quad \text{размером } N_c \times N_c,$$

где I_{r_v} - вектор столбец размерности r_v , состоящий из единиц. Из структуры этой матрицы видно, что

$$SS^T = R^{-1} = (r_i^{-1} \delta_{ij}), \quad i, j = \overline{1, N_c}, \quad (2.16)$$

где R - диагональная матрица весов, выражающая распределения опытов по сериям. Умножим теперь левую и правую часть (2.15) слева на матрицу S и получим

$$\bar{y} = \bar{X}\theta + \varepsilon. \quad (2.17)$$

Легко понять, что вектор $\bar{y}$ в качестве элементов имеет среднее значения отклика по сериям:

$$\bar{y} = Sy = (\bar{y}_1, \bar{y}_2, \dots, \bar{y}_{N_c})^T, \quad \text{причем } \bar{y}_v = \bar{y}(x_v) = r_v^{-1} \sum_{j=1}^{r_v} y_{vj}, \quad v = \overline{1, N_c}.$$

Матрица $\bar{X}$ содержит N_c строк, которые отличаются друг от друга и соответствуют разным сериям:

$$\bar{X} = SX = (f(x_1), f(x_2), \dots, f(x_{N_c}))^T.$$

Вектор усреднённых ошибок имеет вид $\bar{\varepsilon} = (\bar{\varepsilon}_1, \dots, \bar{\varepsilon}_{N_c})^T$, где $\bar{\varepsilon}_v = r_v^{-1} \sum_{j=1}^{r_v} \varepsilon_{vj}$, $v = \overline{1, N_c}$.

Учитывая, что

$$E(\varepsilon) = O_{N_c, 1}; \quad E(\varepsilon \varepsilon^T) = D(\varepsilon) = I_N \sigma^2,$$

Получим

$$E(\bar{\varepsilon}) = O_{N_c, 1}; \quad E(\bar{\varepsilon} \bar{\varepsilon}^T) = D(\bar{\varepsilon}) = R^{-1} \sigma^2 = SI_N \sigma^2 S^T.$$

Сумму взвешенных средних квадратичных отклонений для модели (2.15) можно записать в виде квадратичной формы

$$S(\theta) = \varepsilon^T D^{-1}(\varepsilon) \varepsilon = (y - X\theta)^T \sigma^{-2} (y - X\theta).$$

Опуская сомножитель σ^{-2} , запишем

$$SS(\theta) = (y - X\theta)^T (y - X\theta). \quad (2.18)$$

Для модели (2.17) аналогичная мера рассогласования будет иметь вид

$$S_1(\theta) = \bar{\varepsilon}^T D^{-1}(\bar{\varepsilon})\bar{\varepsilon} = (\bar{y} - \bar{X}\theta)^T R\sigma^{-2}(\bar{y} - \bar{X}\theta).$$

Опуская σ^{-2} , запишем

$$SS_1(\theta) = (\bar{y} - \bar{X}\theta)^T R(\bar{y} - \bar{X}\theta). \quad (2.19)$$

Матрица весов R может быть представлена в виде произведения

$$R = Q^T Q, \quad (2.20)$$

где Q - матрица размером $N \times N_c$

$$Q = \begin{bmatrix} I_{r_1} & . & . & . & 0 \\ . & I_{r_2} & & & . \\ . & & . & & . \\ . & & & . & . \\ 0 & . & . & . & I_{r_{N_c}} \end{bmatrix}, \text{ где } I_{r_v} - \text{ столбец единиц}$$

Подставляя (2.20) в (2.19) получим

$$SS_1(\theta) = (Q\bar{y} - X\theta)^T (Q\bar{y} - X\theta), \quad (2.21)$$

где вектор $Q\bar{y}$ – имеет уже N – составляющих, как и вектор y .

Суммы $SS(\theta)$ и $SS_1(\theta)$ связаны между собой. Используя перегруппировку слагаемых первой суммы, можно показать, что

$$SS(\theta) = SS_1(\theta) + SS_E, \quad (2.22)$$

где

$$SS_E = (y - Q\bar{y})^T (y - Q\bar{y}) \quad (2.23)$$

и не зависима от θ . Поэтому совершенно безразлично какой мерой рассогласования пользоваться для отыскания оценок параметров: (2.18), (2.19) или (2.21).

Точечную оценку дисперсии ошибки эксперимента в случае, когда имеется повторные опыты лучше всего определять через сумму (2.23) по формуле

$\hat{\sigma}_E^2 = SS_E / f_E$, где f_E – число степеней свободы суммы SS_E , равное

$f_E = N - N_c = \sum_{v=1}^{N_c} (r_v - 1)$. Оценка $\hat{\sigma}_E^2$ есть не что иное, как объединенная оценка

выборочных дисперсий ошибки опыта, которые можно найти по результатам отдельных серий повторных опытов. Оценка дисперсии ошибки в v -ой серии повторных опытов равна

$$\hat{\sigma}_{Ev}^2 = \frac{SS_{Ev}}{f_{Ev}}, \text{ где } SS_{Ev} = \sum_{j=1}^{r_v} (y_{vj} - \bar{y}_v)^2, f_{Ev} = r_v - 1.$$

Отсюда следует, что объединенная оценка по всем сериям имеет вид

$$\hat{\sigma}_E^2 = \sum_{v=1}^{N_{\hat{a}}} \frac{f_{Ev}}{f_E} \hat{\sigma}_{Ev}^2 = \sum_{v=1}^{N_{\hat{a}}} SS_{Ev} / f_E = SS_E / f_E.$$

Заметим, что оценка $\hat{\sigma}_E^2$ определяется независимо от модели.

При $\theta = \hat{\theta}$ имеем следующие соотношения:

$$SS_R \equiv SS(\hat{\theta}) = SS_1(\hat{\theta}) + SS_E, \quad (2.24)$$

где SS_R - остаточная сумма со степенями свободы $f_R = N - m$;

$SS_{LF} = SS_1(\hat{\theta})$ - мера степени неадекватности представления экспери

ментальных данных с помощью выбранной модели. Она оценивает рассеивания средних значений отклика по сериям относительно расчетной регрессионной зависимости с учётом весов (**Lack of fit – неадекватность**).

Оценка дисперсии на основе суммы SS_{LF} вычисляется по формуле

$$\hat{\sigma}_{LF}^2 = SS_{LF} / f_{LF}, \quad f_{LF} = N_c - m. \quad (2.25)$$

Оценка величины σ^2 на основе SS_R выражается отношением

$$\hat{\sigma}_R^2 = SS_R / f_R, \quad f_R = N - m.$$

Оценка величины $\hat{\sigma}_R^2$ является, по существу, объединённой. Используя (2.24) и учитывая, что $f_R = f_{LF} + f_E$ можно записать

$$\hat{\sigma}_R^2 = (f_{LF} / f_R) \hat{\sigma}_{LF}^2 + (f_E / f_R) \hat{\sigma}_E^2.$$

Задача

1. Модель $\eta_1 = \theta_1 x + \theta_2 x^2$
2. Гипотеза принимается

Билет №16

Вопрос 1

Часто класс допустимых решений задают в виде параметрического множества функций.

Вопрос 2

Оценки параметров распределены по нормальному закону с мат.ожиданием равным истинным значениям параметров и дисперсией $\sigma^2 (X^T X)^{-1}$

Вопрос 3

$$D(\hat{\theta}) = E(\hat{\theta} - E\hat{\theta})(\hat{\theta} - E\hat{\theta})^T = E[(X^T X)^{-1} X^T \varepsilon \varepsilon^T X (X^T X)^{-1}] = \sigma^2 (X^T X)^{-1} \blacksquare$$

Вопрос 4

Это проверка того, что модель системы не изменилась со временем.

Вопрос 5

Число главных компонент определяется по числу относительно больших собственных значений информационной матрицы

Вопрос 6

$$I - P = I - X(X^T X)^{-1} X^T.$$

$$(I - P)^2 = I - 2P + P^2 =$$

$$= I - 2X(X^T X)^{-1} X^T + X(X^T X)^{-1} X^T X(X^T X(X^T X)^{-1} X^T)^{-1} X^T = I - P$$

Вопрос 7

Для линейной модели с двумя факторами при 4 наблюдениях в углах квадрата информационная матрица будет диагональной с диагональю 4. При 8 наблюдениях (повторы) диагональ будет равна 8 и т.д.

Вопрос 8

Доступный обобщенный МНК. В случае обнаружения эффекта гетероскедастичности можно провести коррекцию оценок параметров модели с помощью взвешенного МНК с весами $\hat{\sigma}_i^{-2} = h^{-1}(z_i^T \hat{\alpha})$. Может оказаться, что для некоторых t $h^{-1}(z_t^T \hat{\alpha}) < 0$. Если число таких наблюдений невелико, то их можно просто выбросить, в противном случае использовать представление $\sigma_i^2 = \exp(h(z_i^T \alpha))$.

Вопрос 9

В качестве критерия величины смещения (функции риска) выберем среднюю сумму квадратов ошибок (ССКО)

$$L_{\tilde{\theta}}(\theta) = E(\tilde{\theta} - \theta)^T (\tilde{\theta} - \theta),$$

где $\tilde{\theta}$ - некоторая оценка параметров $\theta \in R^m$.

Для МНК оценки

$$L_{\hat{\theta}}(\theta) = \sigma^2 \text{tr}(X^T X)^{-1} \rightarrow \infty \text{ при } \lambda_1 \rightarrow 0,$$

т.е. МНК оценка в условиях усиления коллинеарности ($\lambda_1 \rightarrow 0$) не обеспечивает ограниченность ССКО. Это будет происходить и в условиях увеличения числа наблюдений. В нормальных условиях (при отсутствии мультиколлинеарности) $\text{tr}(X^T X)^{-1} \rightarrow 0$ при $N \rightarrow \infty$, т.е. теоретически гарантируется состоятельность в среднеквадратичном.

Вопрос 10

Переменные, которым соответствуют элементы вектора y для минимального собственного числа с значимо отличающимися от 0 значениями и порождают мультиколлинеарность.

Вопрос 11

$$p_{ii} = f^T(x_i)(X^T X)^{-1} f(x_i)$$

. На практике руководствуются следующими правилами. Значения $p_{ii} \leq 0,2$ позволяют считать надежными данные наблюдения, значения $0,2 \leq p_{ii} \leq 0,5$ относят данное наблюдение к классу рискованных, а при значениях $p_{ii} > 0,5$ i -е наблюдение можно принять за точку разбалансировки.

Вопрос 12

В 4 раза.

Вопрос 13

Монотонность коэффициента детерминации от числа включенных в модель регрессоров легко доказывается, если обратиться к его выражению:

$$R^2 = 1 - \frac{RSS}{\sum (y_i - \bar{y})^2}$$

Вопрос 14

При наличии повторной выборки можно предложить критерий селекции, в основу которого положена идея идеального внешнего критерия. При вычислении идеального внешнего критерия вместо зашумленных значений отклика используются точные (незашумленные) значения. Однако в реальной ситуации эти значения нам не известны. В случае повторной выборки вместо неизвестных значений вектора можно использовать и их оценку в виде выборочного среднего $\bar{y}$:

$$\bar{y}^T = (\bar{y}_A^T, \bar{y}_B^T), \bar{y}_{A_i} = (y_{A_i} + y_{B_i})/2, i = 1, \dots, n_A, \bar{y}_B = \bar{y}_A.$$

С использованием вектора $\bar{y}$ можно предложить следующий субидеальный критерий стабильности:

$$\tilde{S}^2 = \|\bar{y}_A - X_A \hat{\theta}_A\|^2 + \|\bar{y}_B - X_B \hat{\theta}_B\|^2.$$

Вопрос 15

С целью управления масштабом оценок введем в рассмотрение функцию стоимости

$$C = \sum_{i=1}^n (y_i - f(x_i)^T \theta)^2 + \sum_{j=1}^m \lambda_j \theta_j^2,$$

где второе слагаемое рассматривается как штраф при условии, что $\lambda_j \geq 0$. Решим данную оптимизационную задачу $C \rightarrow \min$.

Представим целевую функцию в матричном виде

$$C = (y - X\theta)^T (y - X\theta) + \theta^T \Lambda \theta,$$

где $\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_m \end{pmatrix}$.

Найдем решение из условия равенства нулю первых производных:

$$\frac{\partial C}{\partial \theta_j} = 2X^T y + 2X^T X \hat{\theta} + 2\Lambda \hat{\theta} = 0,$$

$$\hat{\theta} = (X^T X + \Lambda)^{-1} X^T y.$$

Ридж-оценки впервые были введены А. Гоэрлом при исследовании регрессии 2-го порядка (ridge – гребень, хребет). Название "ридж-оценки" навеяно следующими наблюдениями. При сильной мультиколлинеарности линии равного уровня доверительного

эллипсоида для разного уровня значимости (доверительной вероятности) становятся сильно вытянутыми и визуально похожими на хребет на географической карте. Данные оценки использовались именно в условиях мультиколлинеарности и поэтому получили название ридж-оценок.

Понятно, что ридж-ценки существенно зависят от выбранной матрицы Λ , в силу чего имеют обозначение

$$\hat{\theta}(\Lambda) = (X^T X + \Lambda)^{-1} X^T y.$$

Часто матрицу Λ задают диагональной в виде $\Lambda_{ii} = \lambda \cdot (X^T X)_{ii}$, $\lambda \geq 0$, т.е. пропорциональной диагональным элементам матрицы $X^T X$. Иногда применяют ещё более простой вариант $\Lambda = \lambda \cdot I_m$. Можно показать, что к данному случаю сводится предыдущий.

Обозначим

$$D = \begin{pmatrix} (X^T X)_{11} & & 0 \\ & \ddots & \\ 0 & & (X^T X)_{mm} \end{pmatrix},$$

$$Z = XD^{-1/2}, \quad \beta = D^{1/2} \theta.$$

Тогда исходная модель

$$y = X\theta + \varepsilon$$

сводится в новых обозначениях к модели

$$y = Z\beta + \varepsilon = XD^{-1/2} D^{1/2} \theta + \varepsilon = X\theta + \varepsilon.$$

Ридж-оценка $\hat{\beta}(\lambda)$ с матрицей $\lambda \cdot I_m$ будет равна

$$\begin{aligned} \hat{\beta}(\lambda) &= (Z^T Z + \lambda I)^{-1} Z^T y = (D^{-1/2} X^T X D^{-1/2} + \lambda I)^{-1} D^{-1/2} X^T y = \\ &= (X^T X + \lambda D)^{-1} D D^{-1/2} X^T y = D^{1/2} (X^T X + \Lambda)^{-1} X^T y, \end{aligned}$$

где $\Lambda_{ii} = \lambda (X^T X)_{ii}$. Поэтому $\hat{\theta}(\Lambda) = D^{-1/2} \hat{\beta}(\lambda)$.

Итак, можно рассматривать более простой случай

$$\hat{\beta}(\lambda) = (Z^T Z + \lambda I)^{-1} Z^T y.$$

Свойства этой оценки:

- 1) $\hat{\beta}(0)$ - МНК-оценка $\hat{\beta}$;
 - 2) $\hat{\beta}(\lambda)$ - является линейным преобразованием МНК-оценки и является смещённой;
- Действительно,

$$Z^T y = (I + \lambda (Z^T Z)^{-1})^{-1} (Z^T Z)^{-1} Z^T y = B \hat{\beta},$$

где $B = (I + \lambda Z^T Z)^{-1}$ - матрица линейного преобразования и $B \neq I$, что приводит к смещению оценок $\hat{\beta}(\lambda)$ в отличие от $\hat{\beta}$, которые несмещены.

3) в классе оценок с фиксированной длиной ридж-оценка минимизирует сумму квадратов отклонений;

4) верно и обратное утверждение: ридж-оценка имеет минимальную длину в классе оценок с заданным значением суммы квадратов отклонений (RSS).

Доказано, что всегда существует такое значение параметра λ , что ССКО для оценки $\hat{\beta}(\lambda)$ будет меньше чем ССКО для оценки $\hat{\beta}(0)$. Показано также, что

$$\hat{\beta}^T(\lambda)\hat{\beta}(\lambda) = (\mu_{\max} / (\mu_{\max} + \lambda)) \hat{\beta}^T \hat{\beta},$$

где $\mu_{\max}$ - максимальное собственное число информационной матрицы $Z^T Z$. Это означает, что длина ридж-оценки будет уменьшаться с увеличением параметра регуляризации, но это и так понятно из рассмотрения целевой функции, где используется штраф за длину оценки.

Задача

1. Модель $\eta_1 = \theta_1 x + \theta_2 x^2$
2. Гипотеза принимается

Билет № 17

Вопрос 1

Нужно отметить, что использование симметричных форм критериев всегда целесообразно: во-первых, снижается чувствительность критерия к способу разделения данных на части, и, во-вторых, усредняется влияние помех, действующих на различных частях выборки.

Вопрос 2

Для того, что бы было возможным обращение информационной матрицы.

Вопрос 3

Для получения несмещенной оценки.

Вопрос 4

Ридж-оценки впервые были введены А. Гоэрлом при исследовании регрессии 2-го порядка (ridge – гребень, хребет). Название "ридж-оценки" навеяно следующими наблюдениями. При сильной мультиколлинеарности линии равного уровня доверительного эллипсоида для разного уровня значимости (доверительной вероятности) становятся сильно вытянутыми и визуально похожими на хребет на географической карте. Данные оценки использовались именно в условиях мультиколлинеарности и поэтому получили название ридж-оценок.

Вопрос 5

Необходимо промасштабировать информационную матрицу на ее след.

Вопрос 6

Условие минимума известно – оно определяется как

$$X^T \underbrace{X\hat{\theta}}_{\hat{y}} = X^T y.$$

Отсюда $X^T(y - X\hat{\theta}) = 0$ и далее $X^T(y - \hat{y}) = 0$, т.е. вектор остатков $\hat{e} = y - \hat{y}$ ортогонален Ω .

Вопрос 7

В 2 раза.

Вопрос 8

Ясно, что в реальных ситуациях редко бывает так, что параметр автокорреляции ρ известен. При неизвестном ρ процедуры оценивания параметров θ носят итеративный характер.

Процедура Кохрейна-Оркатта.

1. Для исходной модели оцениваются $\hat{\theta}$ и $\hat{e}$;
2. Рассматривается регрессия $\hat{e}_t = \rho \hat{e}_{t-1} + v_t$ и вычисляется оценка $\hat{\rho}$;
3. Проводится ОМНК оценивание исходной модели с $\hat{\Omega}^{-1}$, где в качестве ρ используется его оценка $\hat{\rho}$;
4. Вычисляется новый вектор остатков $\hat{e}$;
5. Процедура повторяется с п.2.

Останов производится после нескольких итераций или после того как очередное $\hat{\rho}$ мало отличается от предыдущего.

Процедура Хилдрета-Лу.

По схеме ОМНК производится решение следующей экстремальной задачи:

$$\hat{\rho} = \underset{\rho \in [-1, +1]}{\text{Arg min}} \text{RSS}.$$

Попутно при этом мы получаем оценки для вектора параметров θ .

Вопрос 9

$$\hat{\theta} = (X^T X)^{-1} X^T y. \quad \hat{\theta} = (n)^{-1} X^T y = \sum_{i=1}^n y_i / n$$

Вопрос 10

$$P = X(X^T X)^{-1} X^T. \quad P^2 = X(X^T X)^{-1} X^T X(X^T X(X^T X)^{-1} X^T)^{-1} X^T = P$$

Вопрос 11

Точечную оценку дисперсии ошибки эксперимента в случае, когда имеется повторные опыты лучше всего определять через сумму (2.23) по формуле

$\hat{\sigma}_E^2 = SS_E / f_E$, где f_E число степеней свободы суммы SS_E , равное

$f_E = N - N_c = \sum_{v=1}^{N_c} (r_v - 1)$. Оценка $\hat{\sigma}_E^2$ есть не что иное, как объединенная оценка

выборочных дисперсий ошибки опыта, которые можно найти по результатам отдельных серий повторных опытов. Оценка дисперсии ошибки в v -ой серии повторных опытов равна

$$\hat{\sigma}_{Ev}^2 = \frac{SS_{Ev}}{f_{Ev}}, \quad \text{где } SS_{Ev} = \sum_{j=1}^{r_v} (y_{vj} - \bar{y}_v)^2, \quad f_{Ev} = r_v - 1.$$

Отсюда следует, что объединенная оценка по всем сериям имеет вид

$$\hat{\sigma}_E^2 = \sum_{v=1}^{N_c} \frac{f_{Ev}}{f_E} \hat{\sigma}_{Ev}^2 = \sum_{v=1}^{N_c} SS_{Ev} / f_E = SS_E / f_E.$$

Вопрос 12

. Рассмотрим случай когда в модели присутствует свободный член. В этом случае сумма остатков равно нулю, т.е.

$$\sum (y_i - \hat{y}_i) = 0. \quad (a)$$

Покажем это. Для этого сумму квадратов отклонений для модели со свободным членом

$$\sum (y_i - \theta_0 - \theta_1 x_{1i} - \dots - \theta_{m-1} x_{m-1,i})^2$$

продифференцируем по θ_0 :

$$\sum (y_i - \hat{\theta}_0 - \hat{\theta}_1 x_{1i} - \dots - \hat{\theta}_{m-1} x_{m-1,i}) = 0 \text{ или } \sum (y_i - \hat{y}_i) = 0.$$

Вопрос 13

Для ее вычисления необходимо обратить матрицу коэффициентов корреляций, которая может быть плохо обусловлена в условиях мультиколлинеарности.

Вопрос 14

$$p_{ii} = f^T(x_i)(X^T X)^{-1} f(x_i)$$

. На практике руководствуются следующими правилами. Значения $p_{ii} \leq 0,2$ позволяют считать надежными данные наблюдения, значения $0,2 \leq p_{ii} \leq 0,5$ относят данное наблюдение к классу рискованных, а при значениях $p_{ii} > 0,5$ i -е наблюдение можно принять за точку разбалансировки.

Вопрос 15

Заменим предположение А.2 на более общее предположение

$$A.5 \quad E(\varepsilon \varepsilon^T) = \sigma^2 \Omega = V,$$

где $tr \Omega = n$, Ω - симметричная п.о.

Эффективный оценщик для θ в этом случае можно получить, если исходную модель трансформировать в модель, соответствующую условиям теоремы Гаусса – Маркова.

Поскольку Ω - симметричная п.о., то её можно выразить как PP^T , где P – невырожденная матрица, т.е.

$$\Omega = PP^T \text{ и } \Omega^{-1} = (P^T)^{-1} P^{-1} = (P^{-1})^T P^{-1}.$$

$$(\text{Свойства обратной матрицы } (AB)^{-1} = B^{-1} A^{-1}).$$

Перепишем модель (1.4):

$$y^* = X^* \theta + \varepsilon^*,$$

$$\text{где } y^* = P^{-1} y, \quad X^* = P^{-1} X, \quad \varepsilon^* = P^{-1} \varepsilon.$$

Для данной модели выполняются свойства:

$$E(\varepsilon^*) = P^{-1} E(\varepsilon) = 0, \quad E(\varepsilon^* \varepsilon^{*T}) = \sigma^2 I, \quad rg X^* = rg P^{-1} X = rg X = m.$$

Мы можем теперь формально применить теорему Гаусса-Маркова для вычисления оценок:

$$\begin{aligned}\hat{\theta}_{ОМНК} &= (X^{*T} X^*)^{-1} X^{*T} y^* = (X^T P^{-1T} P^{-1} X)^{-1} X^T P^{-1T} P^{-1} y = \\ &= (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} y = (X^T V^{-1} X)^{-1} X^T V^{-1} y.\end{aligned}\quad (3.1)$$

Данные оценки получили название оценок по обобщенному МНК. Их также можно получить отталкиваясь от нормированной на дисперсионную матрицу сумму квадратов ошибок:

$$\begin{aligned}S(\theta) &= (y - X\theta)^T V^{-1} (y - X\theta) = \\ &= y^T V^{-1} y - 2\theta^T X^T V^{-1} y + \theta^T X^T V^{-1} X \theta.\end{aligned}$$

Первые частные производные $S(\theta)$ по параметрам θ имеют вид

$$\frac{\partial S(\theta)}{\partial \theta} = -2X^T V^{-1} y + 2X^T V^{-1} X \theta.$$

Приравнивая их нулю, получаем систему нормальных уравнений

$$X^T V^{-1} X \hat{\theta}_{ОМНК} = X^T V^{-1} y \text{ с решением } \hat{\theta}_{ОМНК} = (X^T V^{-1} X)^{-1} X^T V^{-1} y.$$

Дисперсионная матрица оценок имеет вид

$$D(\hat{\theta}_{ОМНК}) = \sigma^2 (X^{*T} X^*)^{-1} = \sigma^2 (X^T \Omega^{-1} X)^{-1} = (X^T V^{-1} X)^{-1}.$$

Если в структуре матрицы $E(\varepsilon \varepsilon^T) = \sigma^2 \Omega = V$ неизвестной величиной является дисперсия σ^2 , то ее несмещенной оценкой будет оценка:

$$\hat{\sigma}^2 = \frac{1}{n-m} (Y - X \hat{\theta}_{ОМНК})^T \Omega^{-1} (Y - X \hat{\theta}_{ОМНК}).$$

Задача

1. Модель $\eta_1 = \theta_1 + \theta_2 x$
2. Гипотеза принимается

Билет № 18

Вопрос 1

Термин "сферичность" возмущений происходит из того наблюдения, что в многомерном пространстве распределения помехи ε при $Var(\varepsilon) = \sigma^2 I_n$ (случай равноточных наблюдений) точки с одинаковой вероятностью образуют сферы. В рассматриваемом случае неравноточных наблюдений вместо сфер будут образовываться многомерные эллипсоиды.

Вопрос 2

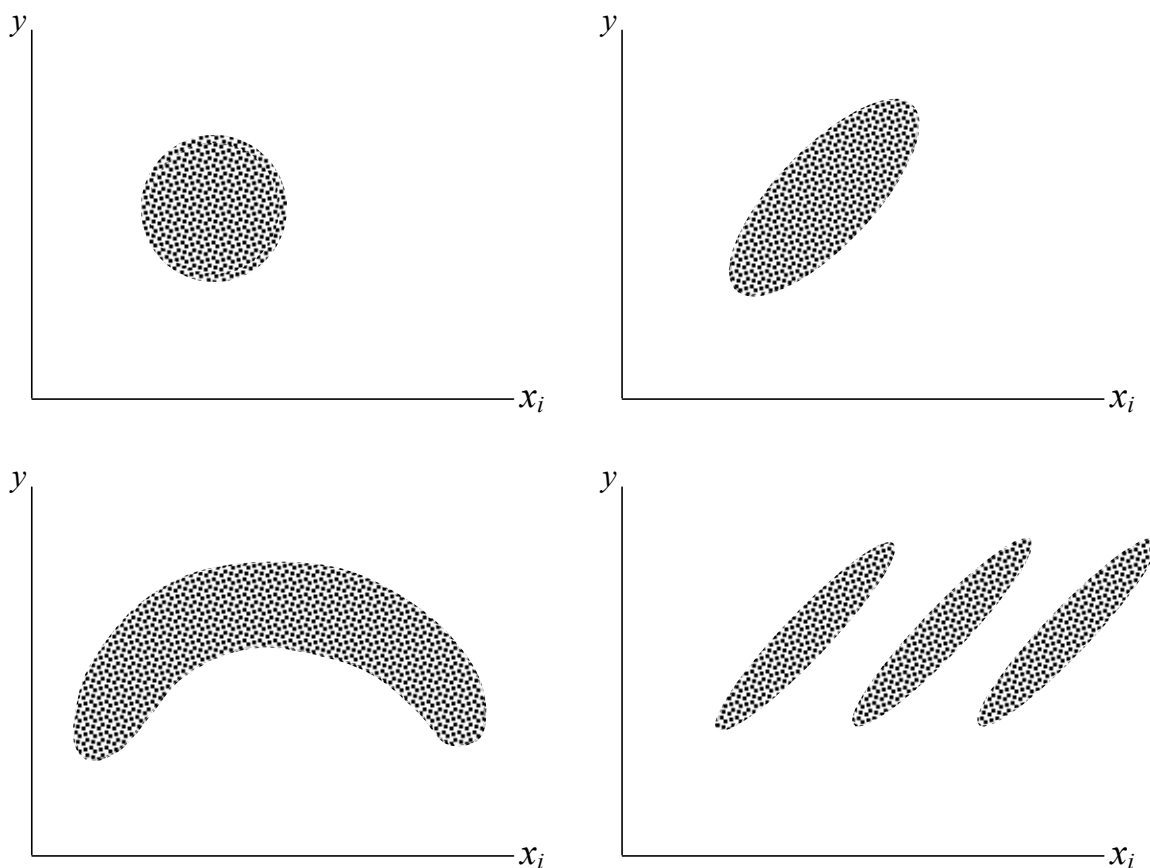
Нет.

Вопрос 3

При предварительном анализе характера исследуемых зависимостей ограничиваются в основном изучением “геометрии” парных корреляционных полей. Под корреляционным полем переменных (u, v) понимается графическое представление имеющихся измерений $(u_1, v_1), (u_2, v_2), \dots, (u_n, v_n)$ этих переменных в плоскости (u, v) .

Анализ парных корреляционных полей состоит обычно в следующем:

- а) построение на основании имеющихся исходных данных корреляционных полей для всевозможных пар переменных вида (x_j, x_l) и (x_j, y) — всего таких пар будет $k(k+1)/2$. Этот процесс легко автоматизируется с помощью средств современных ЭВМ;
- б) визуальное прослеживание характера вытянутости каждого корреляционного поля (эллипсоидально-линейное, нелинейное-монотонное, с наличием экстремумов и т. п.);



Это примеры корреляционных полей для пассивного эксперимента.

Вопрос 4

Выделяющиеся точки графика: некоторые из остатков могут по абсолютной величине сильно превосходить все остальные остатки. Если максимальное значение остатка $|d_{\max}| > \delta$, где δ - некоторое заранее выбранное число, то наблюдение, имеющее такой остаток, является аномальным. Величина остатка аномального наблюдения зависит от объема выборки n : при $n > 20$ его можно обнаружить по правилу «трех сигм» ($\delta = 3$), при $n < 20$, например, по критерию Смирнова - Граббса. В этом случае вычисляемая статистика имеет вид

$$D = (d_{\max} - \bar{d}) / S_d, \text{ где } \bar{d} - \text{среднее значение остатков, } S_d^2 = 1/n \left(\sum_{i=1}^n (d_i - \bar{d})^2 \right).$$

Вопрос 5

Свойство 4, поскольку преследуется цель уменьшения нормы оценок параметров.

Вопрос 6

Доступный обобщенный МНК. В случае обнаружения эффекта гетероскедастичности можно провести коррекцию оценок параметров модели с помощью взвешенного МНК с весами $\hat{\sigma}_i^{-2} = h^{-1}(z_i^T \hat{\alpha})$. Может оказаться, что для некоторых i $h^{-1}(z_i^T \hat{\alpha}) < 0$. Если число таких наблюдений невелико, то их можно просто выбросить, в противном случае использовать представление $\sigma_i^2 = \exp(h(z_i^T \alpha))$.

Вопрос 7

Для критерия Δ_{abs}^2 , например, его среднее значение будет определяться величиной

$$\bar{\Delta}_{abs}^2 = u^T D_{\Delta_{abs}(p)} u + \sigma^2 \cdot p.$$

С увеличением σ^2 минимум критерия в его зависимости от сложности модели смещается влево, т.е. выбираются все более простые модели. Абсолютно помехоустойчивый критерий получил свое название по причине того, что второе слагаемое для него менее всего возрастает при увеличении дисперсии помехи. Другими словами, при увеличении дисперсии помехи минимум данного критерия с наименьшей скоростью смещается влево в сторону все более простых моделей.

Вопрос 8

Алгоритм состоит в следующем.

На первом шаге выбирается некоторое число лучших моделей, состоящих из одного регрессора. На втором этапе к этим моделям добавляются различные регрессоры по одному, и отбираются модели с двумя аргументами, улучшающие величину критерия, и т.д. до его ухудшения или до построения полной модели.

На последнем шаге можно из оставшихся лучших моделей выбрать одну по другому критерию.

Вопрос 9

$$D(\hat{\theta}) = E(\hat{\theta} - E\hat{\theta})(\hat{\theta} - E\hat{\theta})^T = E[(X^T X)^{-1} X^T \varepsilon \varepsilon^T X (X^T X)^{-1}] = \sigma^2 (X^T X)^{-1} \blacksquare$$

Вопрос 10

Имеем $X^T (y - X\hat{\theta}) = 0$ и далее $X^T (y - \hat{y}) = 0$, т.е. вектор остатков $\hat{e} = y - \hat{y}$ ортогонален Ω . В тоже время справедливо представление $y = \hat{y} + \hat{e}$. Становится понятным, что $\hat{y}$ есть проекция y на $R(X)$, а именно $\hat{y} = Py$. Можно рассматривать также подпространство

$R^\perp(X)$ – ортогональное дополнение для $R(X)$ в $R(X)$. Поскольку

$$\hat{e} = y - \hat{y} = y - X\hat{\theta} = (I_n - X(X^T X)^{-1} X^T) y = (I_n - P)y,$$

то $M = I_n - P$ – есть матрица проектирования вектора y на подпространство $R^\perp(X)$.

Вопрос 11

$$\hat{\theta} = (X^T X)^{-1} X^T y. \quad \hat{\theta} = (n)^{-1} X^T y = \sum_{i=1}^n y_i / n$$

Вопрос 12

В качестве критерия величины смещения (функции риска) выберем среднюю сумму квадратов ошибок (ССКО)

$$L_{\tilde{\theta}}(\theta) = E(\tilde{\theta} - \theta)^T (\tilde{\theta} - \theta),$$

где $\tilde{\theta}$ - некоторая оценка параметров $\theta \in R^m$.

Для МНК оценки

$$L_{\hat{\theta}}(\theta) = \sigma^2 \text{tr}(X^T X)^{-1} \rightarrow \infty \text{ при } \lambda_1 \rightarrow 0,$$

т.е. МНК оценка в условиях усиления коллинеарности ($\lambda_1 \rightarrow 0$) не обеспечивает ограниченность ССКО. Это будет происходить и в условиях увеличения числа наблюдений. В нормальных условиях (при отсутствии мультиколлинеарности) $\text{tr}(X^T X)^{-1} \rightarrow 0$ при $N \rightarrow \infty$, т.е. теоретически гарантируется состоятельность в среднеквадратичном.

Вопрос 13

Переменные, которым соответствуют элементы вектора $\underline{y}$ для минимального собственного числа с значимо отличающимися от 0 значениями и порождают мультиколлинеарность.

Вопрос 14

Статистика всегда удовлетворяет неравенству

$$0 < DW < 4.$$

При отсутствии автокорреляции статистика DW близка к двум, при положительной автокорреляции она близка к нулю, а при отрицательной к 4. Точное распределение статистики DW зависит не только от числа наблюдений n и количества регрессоров m , но и от всей матрицы X , что означает невозможность ее табулирования. Тем не менее, Дарбин и Утсон показали, что есть определенные границы, зависящие только от числа наблюдений n , количества регрессоров m и уровня значимости (поэтому могут быть затабулированы). Нулевая гипотеза состоит в том, что $H_0: \rho = 0$, против альтернативы $H_1: \rho > 0$ или $H_1': \rho < 0$. Обозначим d_α^U , d_α^L – нижний и верхний уровни для доверительной границы, которые не зависят от X . Влияние неопределенности в матрице X все же привело к появлению зоны неопределенности в принятии или отвержении первоначальной гипотезы. Результаты можно представить

$4 - d_\alpha^L < DW < 4$	гипотеза $H_0: \rho = 0$ отвергается, есть отрицательная корреляция;
$4 - d_\alpha^U < DW < 4 - d_\alpha^L$	неопределенность;
$d_\alpha^U < DW < 4 - d_\alpha^U$	$H_0: \rho = 0$ не отвергается;
$d_\alpha^L < DW < d_\alpha^U$	неопределенность;
$0 < DW < d_\alpha^L$	$H_0: \rho = 0$ отвергается, есть положительная корреляция.

Вопрос 15

Модифицированные жордановы исключения или выметания

Говорят, что матрица $A = [(a_{ij})]$, $i, j = \overline{1, m}$ выметена по k -ой строке и k -му столбцу (или по k -му ведущему элементу), если она преобразована к матрице $A^* = [(a_{ij}^*)]$, у которой

$$a_{kk}^* = \frac{1}{a_{kk}}, \quad a_{ik}^* = \frac{-a_{ik}}{a_{kk}} \quad (i \neq k),$$

$$a_{kj}^* = \frac{a_{kj}}{a_{kk}} \quad (j \neq k), \quad a_{ij}^* = a_{ij} - \frac{a_{ik}a_{kj}}{a_{kk}} \quad (i, j \neq k).$$

Оператор выметания обладает следующими свойствами:

- 7) оператор выметания обратим, так что двукратное применение выметания по одному и тому же (в смысле номера) ведущему элементу матрицы равносильно неприменению этого оператора.
- 8) Выметание коммутативно, так что выметание в порядке k , l -ых элементов равносильно выметанию, проведенному в порядке l , k .

Сформируем матрицу A следующего вида

$$A = \begin{matrix} m & m+1 \\ \begin{pmatrix} X^T X & X^T y \\ y^T X & y^T y \end{pmatrix} \end{matrix} \quad (4.43)$$

Если матрицу A "выметим" по ее первым m ведущим элементам, то на ее месте получим (при условии, что $(X^T X)^{-1}$ существует)

$$B = \left(\begin{array}{c|c} (X^T X)^{-1} & (X^T X)^{-1} X^T y \\ \hline -y^T X (X^T X)^{-1} & y^T y - y^T X (X^T X)^{-1} X^T y \end{array} \right) = \left(\begin{array}{cc} (X^T X)^{-1} & \hat{\theta} \\ -\hat{\theta}^T & RSS \end{array} \right).$$

Если провести выметание по первым i ведущим элементам ($i \leq m$), то в $m+1$ -ом столбце на первых i позициях будут находиться $\hat{\theta}_j$, $j=1,2,\dots,i$, $i \times i$ -матрица будет соответствовать $(X_{(i)}^T X_{(i)})^{-1}$ – главному минору. Кроме того, в силу обратимости и коммутативности выметание выполнение выметания по некоторому ведущему элементу матрицы A приводит либо к включению в модель соответствующего регрессора, либо к исключению этого регрессора из модели, если он в ней уже присутствовал.

Порядок построения моделей.

Используя операцию выметания можно построить эффективную процедуру построения всех возможных моделей.

Рассмотрим простой пример. Пусть

$$f(x, \theta) = \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3, \quad m = 3.$$

Всего возможно $2^3 - 1 = 7$ моделей регрессии

№	№ выметаемого элемента	Модель
1	1	$y_1 = \theta_1 x_1$
2	2	$y_2 = \theta_1 x_1 + \theta_2 x_2$
3	1	$y_3 = \theta_2 x_2$
4	3	$y_4 = \theta_2 x_2 + \theta_3 x_3$
5	1	$y_5 = \theta_2 x_2 + \theta_3 x_3 + \theta_1 x_1$
6	2	$y_6 = \theta_3 x_3 + \theta_1 x_1$

7	1	$y_7 = \theta_3 x_3$
---	---	----------------------

Если $m=4$, то оптимальная последовательность будет следующей

$$S_4 = 1 \ 2 \ 1 \ 3 \ 1 \ 2 \ 1 \ 4 \ 1 \ 2 \ 1 \ 3 \ 1 \ 2 \ 1.$$

Применяя метод индукции, получим

$$S_m = S_{m-1} \ m \ S_{m-1}.$$

Более экономных последовательностей не существует, поскольку здесь каждое выметание порождает новую модель. Всего нужно произвести $2^m - 1$ выметания, чтобы последовательно построить $2^m - 1$ модель.

Последовательность S_m по имени автора получила название счетчик Гарсайда.

Задача

1. Модель $\eta_1 = \theta_1 + \theta_2 x$
2. Гипотеза отвергается

Билет №19

Вопрос 1

Тун 1. Установление самого факта наличия (или отсутствия) статистически значимой связи между откликом и факторами.

Выбор формы связи (т.е. класса допустимых решений F и конкретного вида $\bar{f}(\bar{x}) \in F$) и состава предикторов играет второстепенную роль.

Тун 2. Прогноз (восстановление) неизвестных значений интересующих нас результирующих показателей по заданным значениям факторов x .

Тун 3. Выявление причинных связей между откликами и факторами, частичное управление значениями системы путем регулирования величин входных факторов.

Здесь на первый план выходят задача правильного определения структуры модели.

Вопрос 2

Этап 4 (определение класса допустимых решений).

Главной целью является определение общего вида, структуры искомой связи между $\bar{y}$ и $\bar{x}$, или, другими словами, описание класса функций F , в рамках которого будет производиться поиск конкретного вида интересующей зависимости.

Чаще всего это описание дается в форме некоторого параметрического семейства функций $\bar{f}(\bar{x}, \bar{\theta})$, поэтому этап этот называют также этапом параметризации модели.

Вопрос 3.

Линейные оценки представимы в виде

$$\hat{\theta} = T y,$$

где $y^T = (y_1, \dots, y_n)$, T – матрица $m \times n$.

Вопрос 4

Достоинство – свободно от влияния масштаба. Недостаток – выявляет только парную коллинеарность.

Вопрос 5

Определитель информационной матрицы, минимальное собственное число, мера обусловленности по Нейману-Голдстейну, максимальная парная сопряженность, максимальная сопряженность.

Вопрос 6

Число главных компонент и определяется по числу относительно больших собственных значений матрицы $X^{*T} X^*$. Недобор суммарной изменчивости может при этом составлять несколько процентов.

Вопрос 7

Критерий называется внешним, если его определение основано на использовании новой информации, «свежих» наблюдениях, не использованных при синтезе модели.

Вопрос 8.

Возможно использование критерия скользящего контроля и критерия регулярности.

Вопрос 9

$$D(\hat{\theta}) = E(\hat{\theta} - E\hat{\theta})(\hat{\theta} - E\hat{\theta})^T = E[(X^T X)^{-1} X^T \varepsilon \varepsilon^T X (X^T X)^{-1}] = \sigma^2 (X^T X)^{-1} \blacksquare$$

Вопрос 10

Для получения ММП оценок для Θ и σ^2 необходимо совместно решить

$$\begin{aligned} \frac{\partial \log(L)}{\partial \Theta} &= -\frac{1}{\sigma^2} X^T (y - X\Theta) = 0 \\ \frac{\partial \log(L)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (y - X\Theta)^T (y - X\Theta) = 0 \end{aligned}$$

Решения:

$$\hat{\Theta}_{MML} = (X^T X)^{-1} X^T y, \quad \hat{\sigma}_{MML}^2 = (y - X\hat{\Theta}_{MML})^T (y - X\hat{\Theta}_{MML}) / n.$$

Вопрос 11

Наиболее известный критерий для выявления автокорреляции 1-го порядка – это критерий Дарбина-Уотсона. Для регрессии

$y = X\theta + \varepsilon$ вычисляются остатки $\hat{e} = y - X\hat{\theta}$. По ним вычисляется статистика этого критерия:

$$DW = \sum_{t=2}^n (\hat{e}_t - \hat{e}_{t-1})^2 / \sum_{t=1}^n \hat{e}_t^2.$$

Показано, что

$$DW \cong 2(1 - \hat{\rho}),$$

где $\hat{\rho}$ – коэффициент автокорреляции в регрессии $\hat{e}_t = \rho \hat{e}_{t-1} + v$:

$$\hat{\rho} = \sum_{t=2}^n \hat{e}_t \hat{e}_{t-1} / \sum_{t=2}^n \hat{e}_t^2.$$

Статистика всегда удовлетворяет неравенству

$$0 < DW < 4.$$

При отсутствии автокорреляции статистика DW близка к двум, при положительной автокорреляции она близка к нулю, а при отрицательной к 4.

Вопрос 12

Частный случай 1: проверка равенства отдельного параметра заданной величине c .

$H_0: \theta_j = c$ или $H_0: a^T \theta = c$, где a^T - вектор-строка, в которой на j -ом месте стоит 1, на остальных местах – нули. При $c = 0$ проверяется гипотеза о незначимости параметра.

Обозначим $d_{jj} = ((X^T X)^{-1})_{jj}$ - j -ый диагональный элемент. Тогда

$$F = \frac{(a^T \hat{\theta} - c)^T (a^T \hat{\theta} - c)}{\hat{\sigma}^2 d_{jj}} = \frac{(\hat{\theta}_j - c)^2}{\hat{\sigma}^2 d_{jj}}$$

имеет при гипотезе H_0 распределение $F_{1, n-m}$. т.е. если $F \leq F_{\alpha, 1, n-m}$, где $F_{\alpha, 1, n-m}$ - критическая точка, то гипотеза о незначимости не отклоняется.

Вопрос 13

Главные компоненты должны быть ортогональными, суммарная их изменчивость должна быть равна таковой для исходных факторов, сами они определяются как линейные комбинации исходных факторов.

Вопрос 14

Оператор выметания обладает следующими свойствами:

- 9) оператор выметания обратим, так что двукратное применение выметания по одному и тому же (в смысле номера) ведущему элементу матрицы равносильно неприменению этого оператора.
- 10) Выметание коммутативно, так что выметание в порядке k , l -ых элементов равносильно выметанию, проведенному в порядке l , k .

Если $m=4$, то оптимальная последовательность будет (счетчик Гарсайда) следующей

$$S_4 = 1 \ 2 \ 1 \ 3 \ 1 \ 2 \ 1 \ 4 \ 1 \ 2 \ 1 \ 3 \ 1 \ 2 \ 1.$$

Применяя метод индукции, получим

$$S_m = S_{m-1} \ m \ S_{m-1}.$$

Вопрос 15

Мэлоу (Mallows) предложил использовать в качестве оценки для Δ_p статистику

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} + 2p - n,$$

где $\hat{\sigma}^2$ - подходящая оценка для σ^2 , $RSS_p = (y - \hat{y}_p)^T (y - \hat{y}_p)$ - остаточная сумма квадратов.

Часто берут $\hat{\sigma}^2 = \frac{RSS}{(n-m)}$, где m - полное число регрессоров.

Полагая $\hat{\sigma}^2 \approx \sigma^2$, имеем

$$E[C_p] \cong \frac{1}{\sigma^2} E[RSS_p] + 2p - n,$$

далее

$$E[RSS_p] = E[(y - \hat{y}_p)^T (y - \hat{y}_p)] = E[y^T (I_n - P_p) y] =$$

$$= \sigma^2 \text{tr}[I_n - P_p] + y_{cp}^T (I_n - P_p) y_{cp} = \sigma^2 (n - p) + SSB_p.$$

Поэтому

$$E[C_p] \approx (n - p) + \frac{SSB_p}{\sigma^2} + 2p - n = p + \frac{SSB_p}{\sigma^2} = \Delta_p,$$

т.е. C_p является приблизительно несмещённой оценкой для Δ_p .

Задача

1. Модель $\eta_1 = \theta_1 + \theta_2 x$
2. Гипотеза принимается

Билет № 20

Вопрос 1

Одномерный анализ:

Сравнение математических ожиданий;
Сравнение дисперсий;
Оценивание параметров распределений;
Установление закона распределения;
Отбраковка данных.

Многомерный анализ:

Исследование зависимостей между признаками;
Классификация объектов;
Снижение размерности пространства признаков.

Вопрос 2

Функция $\eta(x, \theta)$ линейна по параметрам, если

$$E(y/x) = \eta(x, \theta) = \theta^T f(x),$$

Вопрос 3

Несмещенность, состоятельность, эффективность.

Вопрос 4

Использование симметричных форм критериев всегда целесообразно: во-первых, снижается чувствительность критерия к способу разделения данных на части, и, во-вторых, усредняется влияние помех, действующих на различных частях выборки.

Вопрос 5

Максимальная парная сопряженность, максимальная сопряженность

Вопрос 6

При предварительном анализе характера исследуемых зависимостей ограничиваются в основном изучением “геометрии” парных корреляционных полей. Под корреляционным полем переменных (u, v) понимается графическое представление имеющихся измерений $(u_1, v_1), (u_2, v_2), \dots, (u_n, v_n)$ этих переменных в плоскости (u, v) .

В результате такого анализа обычно получают формулировку нескольких рабочих гипотез об общем виде искомой зависимости.

Вопрос 7

Частный случай 2. Проверка незначимости уравнения регрессии.

Пусть задана линейная модель наблюдения:

$$y_i = \Theta_0 + \Theta_1 x_{1,i} + \dots + \Theta_{m-1} x_{m-1,i} + \varepsilon_i, \quad i = \overline{1, n}$$

и требуется установить является ли регрессия с заданными регрессорами незначимой, т.е. мы хотим проверить гипотезу $H: \Theta_1 = \Theta_2 = \dots = \Theta_{m-1} = 0$ или другими словами сравнить данную модель с моделью среднего $y = \Theta_0$.

Гипотеза H имеет вид: $A\Theta = 0$, где $A = [0 \mid I_{m-1}]$ - $(m-1) \times m$ матрица ранга $m-1$.

Применима общая теория с $q = m-1$, $RSS = y^T y - \hat{\Theta}^T X^T y$ и

$$RSS_H = \min_{\Theta_0} \sum_{i=1}^n (y_i - \Theta_0)^2 = \sum_{i=1}^n (y_i - \bar{y})^2 = y^T y - n\bar{y}^2.$$

Вопрос 8

Data Mining (или Knowledge Discovery In Data – обнаружение знаний в данных) – исследование и обнаружение машиной в сырых данных скрытых знаний, которые ранее были не известны, нетривиальны, практически полезны, доступны для интерпретации человеком.

Данные: цифровые данные, текст, изображение, видео, звук.

Методы Data Mining находятся на стыке различных направлений информационных технологий: прикладной статистики, баз данных, нейронных сетей, нечетких множеств, генетических алгоритмов, искусственного интеллекта и др.

Основные задачи **Data Mining**: классификация, построение зависимостей, кластеризация, ассоциативные правила, визуализация, поиск исключений.

Вопрос 9

Для линейной модели с двумя факторами при 4 наблюдениях в углах квадрата информационная матрица будет диагональной с диагональю 4. При 8 наблюдениях (повторы) диагональ будет равна 8 и т.д.

Вопрос 10

Условие минимума известно – оно определяется как

$$X^T \underbrace{X \hat{\theta}}_{\hat{y}} = X^T y.$$

Отсюда $X^T (y - X \hat{\theta}) = 0$ и далее $X^T (y - \hat{y}) = 0$, т.е. вектор остатков $\hat{e} = y - \hat{y}$ ортогонален Ω .

Вопрос 11

Поскольку ε_{t-1} и v_t независимы, то

$$E(\varepsilon_t^2) = E(\rho \varepsilon_{t-1} + v_t)^2 = \rho^2 E(\varepsilon_{t-1}^2) + E v_t^2 = \rho^2 E(\varepsilon_{t-1}^2) + \sigma_v^2.$$

Будем предполагать, что ошибки наблюдения ε_t имеют одинаковую дисперсию, т.е. .

$E(\varepsilon_t^2) = E(\varepsilon_{t-1}^2)$. С учетом этого можно получить

$$\sigma_\varepsilon^2 = E(\varepsilon^2) = E(\varepsilon_t^2) = E(\varepsilon_0^2) = \sigma_v^2 / (1 - \rho^2).$$

Относительно ковариации можно получить, умножая (3.6) на ε_{t-1} :

$$E(\varepsilon_t \varepsilon_{t-1}) \equiv \text{cov}(\varepsilon_t \varepsilon_{t-1}) = \rho \sigma_\varepsilon^2$$

и далее $\text{cov}(\varepsilon_t \varepsilon_{t-2}) = \rho^2 \sigma_\varepsilon^2, \dots, \text{cov}(\varepsilon_t \varepsilon_{t-m}) = \rho^m \sigma_\varepsilon^2$.

Таким образом, можно окончательно выписать ковариационную матрицу для ε :

$$V = \frac{\sigma_v^2}{1-\rho^2} \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ . & . & . & \dots & . \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{pmatrix} = \sigma_v^2 \Omega.$$

Вопрос 12

Переменные, которым соответствуют элементы вектора $\underline{y}$ для минимального собственного числа с значимо отличающимися от 0 значениями и порождают мультиколлинеарность.

Вопрос 13

Монотонность коэффициента детерминации от числа включенных в модель регрессоров легко доказывается, если обратиться к его выражению:

$$R^2 = 1 - \frac{RSS}{\sum (y_i - \bar{y})^2}$$

Вопрос 14

При наличии повторной выборки можно предложить критерий селекции, в основу которого положена идея идеального внешнего критерия. При вычислении идеального внешнего критерия вместо зашумленных значений отклика используются точные (незашумленные) значения. Однако в реальной ситуации эти значения нам не известны. В случае повторной выборки вместо неизвестных значений вектора $\underline{y}$ можно использовать и их оценку в виде выборочного среднего $\bar{y}$:

$$\bar{y}^T = (\bar{y}_A^T, \bar{y}_B^T), \bar{y}_{A_i} = (y_{A_i} + y_{B_i})/2, i = 1, \dots, n_A, \bar{y}_B = \bar{y}_A.$$

С использованием вектора $\bar{y}$ можно предложить следующий субидеальный критерий стабильности:

$$\tilde{S}^2 = \|\bar{y}_A - X_A \hat{\theta}_A\|^2 + \|\bar{y}_B - X_B \hat{\theta}_B\|^2.$$

Вопрос 15

Пусть $\tilde{\theta} = C^T y$ - произвольная несмещенная оценка, т.е.

$$E\tilde{\theta} = E(C^T y) = EC^T(X\theta + \varepsilon) = C^T X\theta = \theta.$$

Таким образом для C должно выполняться

$$C^T X = I. \tag{1.6}$$

Обозначим $H = C^T - (X^T X)^{-1} X^T$. Тогда из (1.6)

$$HX = (C^T - (X^T X)^{-1} X^T)X = C^T X - (X^T X)^{-1} X^T X = C^T X - I = 0,$$

т.е.

$$HX = 0.$$

Определим смещение оценки $\tilde{\theta}$:

$$\tilde{\theta} - \theta = C^T y - \theta = C^T X\theta + C^T \varepsilon - \theta = C^T \varepsilon.$$

Вычислим матрицу ковариаций для оценки $\tilde{\theta}$:

$$\text{Cov}(\tilde{\theta}) \equiv D(\tilde{\theta}) = E(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^T = E(C^T \varepsilon \varepsilon^T C) = \sigma^2 C^T C =$$

$$\begin{aligned}
&= \sigma^2 [H + (X^T X)^{-1} X^T] [H^T + X (X^T X)^{-1}] = \\
&= \sigma^2 [HH^T + HX(X^T X)^{-1} + (X^T X)^{-1} X^T H^T + (X^T X)^{-1} X^T X (X^T X)^{-1}] = \\
&= \sigma^2 HH^T + \sigma^2 (X^T X)^{-1} = \sigma^2 HH^T + D(\hat{\theta})
\end{aligned}$$

Действительно второе слагаемое в полученной сумме есть дисперсия оценки $\hat{\theta}$, поскольку: $\hat{\theta} - \theta = (X^T X)^{-1} X^T \varepsilon$ и

$$D(\hat{\theta}) = E(\hat{\theta} - E\hat{\theta})(\hat{\theta} - E\hat{\theta})^T = E[(X^T X)^{-1} X^T \varepsilon \varepsilon^T X (X^T X)^{-1}] = \sigma^2 (X^T X)^{-1} \blacksquare$$

В силу своего представления матрица HH^T - п.п.о., поэтому $D(\tilde{\theta}) \geq D(\hat{\theta}) \blacksquare$

Для того чтобы $HH^T = 0$ необходимо чтобы $H = 0$. Отсюда получаем, что

$$C^T = (X^T X)^{-1} X^T \text{ и } \tilde{\theta} = (X^T X)^{-1} X^T y = \hat{\theta} \blacksquare$$

Задача

1. Модель $\eta_1 = \theta_1 + \theta_2 x$
2. Гипотеза отвергается

Билет №21

Вопрос 1

Методы статистического исследования зависимостей составляют содержание отдельных частей многомерного статистического анализа.

Они формировались и развивались с учетом специфики анализируемых моделей, обусловленной природой изучаемых переменных.

Соответствующая спецификация этих разделов отражена в следующей таблице.

№	Природа откликов	Природа факторов	Название раздела многомерного статистического анализа
1.	Количественная	Количественная	Регрессионный анализ
2.	Количественная	Одна переменная – время	Анализ временных рядов
3.	Количественная	Качественная	Дисперсионный анализ
4.	Количественная	Смешанная	Ковариационный анализ
5.	Неколичественная	Неколичественная	Анализ ранговых корреляций
6.	Неколичественная или отсутствие	Количественная	Дисперсионный анализ, кластерный анализ
7.	Смешанная	Смешанная	Аппарат логических решающих функций

Вопрос 2

Термины «регрессионная модель», «регрессионный анализ» – устоявшиеся, хотя и не совсем удачные.

Термин «регрессия» ввел антрополог Гальтон (1822-1911) в 1885г. Он установил «закон регрессии». Имеется в виду, что «возврат» (регрессия) потомков в отношении их индивидуальных признаков к родителям происходит в «среднем». В частности он рассматривал размер семян и рост детей и родителей. У рослых родителей не обязательно еще более рослые дети и наоборот, у малорослых родителей – еще более малорослые. Он сравнивал средний рост детей Y со средним ростом их родителей X . Гальтон заметил, что

рост детей у высоких (низких) родителей обычно также выше (ниже) среднего роста популяции $\mu \cong \bar{X} \cong \bar{Y}$, но при этом отклонение от μ у детей меньше, чем у родителей. Другими словами, экстремумы в следующем поколении сглаживаются, происходит возвращение назад (*регрессия*) к середине. Гальтон показал, что зависимость Y от X хорошо выражается уравнением $Y - \bar{Y} = (2/3)(X - \bar{X})$.

Вопрос 3

A1: $E(\varepsilon) = 0$; A2: $E(\varepsilon\varepsilon^T) = \sigma^2 I_n$; A3: $rg(X) = m$.

Вопрос 4

Термин "сферичность" возмущений происходит из того наблюдения, что в многомерном пространстве распределения помехи ε при $Var(\varepsilon) = \sigma^2 I_n$ (случай равнооточных наблюдений) точки с одинаковой вероятностью образуют сферы. В рассматриваемом случае неравнооточных наблюдений вместо сфер будут образовываться многомерные эллипсоиды.

Вопрос 5

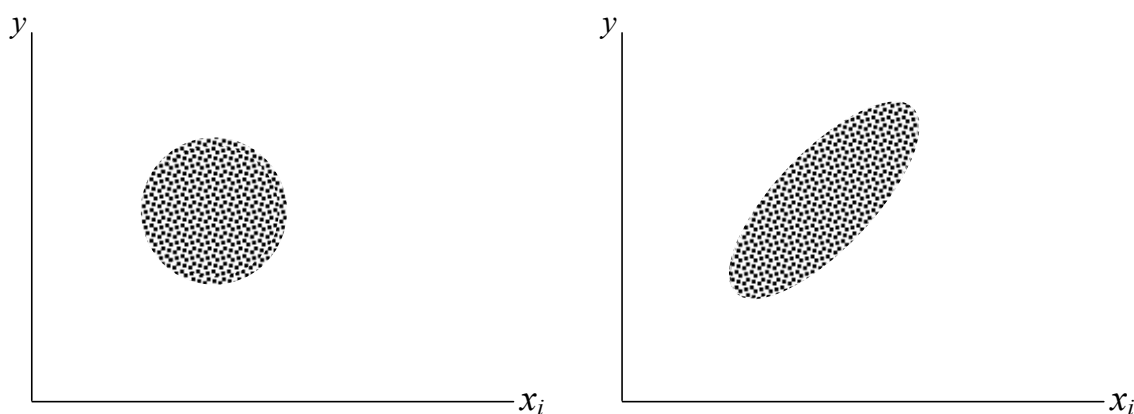
Нет.

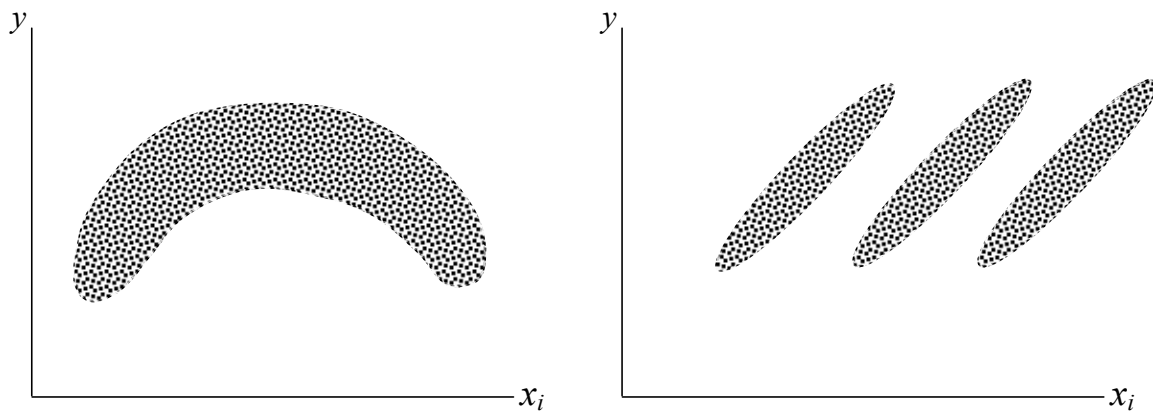
Вопрос 6

При предварительном анализе характера исследуемых зависимостей ограничиваются в основном изучением "геометрии" парных корреляционных полей. Под корреляционным полем переменных (u, v) понимается графическое представление имеющихся измерений $(u_1, v_1), (u_2, v_2), \dots, (u_n, v_n)$ этих переменных в плоскости (u, v) .

Анализ парных корреляционных полей состоит обычно в следующем:

- а) построение на основании имеющихся исходных данных корреляционных полей для всевозможных пар переменных вида (x_j, x_l) и (x_j, y) — всего таких пар будет $k(k+1)/2$. Этот процесс легко автоматизируется с помощью средств современных ЭВМ;
- б) визуальное прослеживание характера вытянутости каждого корреляционного поля (эллипсоидально-линейное, нелинейное-монотонное, с наличием экстремумов и т. п.);





Это примеры корреляционных полей для пассивного эксперимента.

Вопрос 7

Критерий стабильности, регулярности.

Вопрос 8

Этапы: постановочный. Информационный, корреляционный анализ, определение класса допустимых решений, анализ мультиколлинеарности, оценка параметров, проверка адекватности

Вопрос 9

Имеем

$$\hat{\theta} = (X^T X)^{-1} X^T (X\theta + \varepsilon) = \theta + (X^T X)^{-1} X^T \varepsilon;$$

$$E(\hat{\theta}) = E(\theta) + (X^T X)^{-1} X^T E(\varepsilon) = \theta$$

поскольку действует предположение А1.

Вопрос 10

Поскольку $\hat{y}$ и $\hat{e}$ из ортогональных подпространств, то ясно, что $\hat{y}^T \hat{e} = 0$. Далее, поскольку $y = \hat{y} + \hat{e}$, то $y^T y = \hat{y}^T \hat{y} + \hat{e}^T \hat{e}$.

Вопрос 11

Доступный обобщенный МНК. В случае обнаружения эффекта гетероскедастичности можно провести коррекцию оценок параметров модели с помощью взвешенного МНК с весами $\hat{\sigma}_t^{-2} = h^{-1}(z_t^T \hat{\alpha})$. Может оказаться, что для некоторых t $h^{-1}(z_t^T \hat{\alpha}) < 0$. Если число таких наблюдений невелико, то их можно просто выбросить, в противном случае использовать представление $\sigma_t^2 = \exp(h(z_t^T \alpha))$.

Вопрос 12

Можно рекомендовать использовать информационную матрицу, нормированную на след:

$$X^T X / \text{tr} X^T X.$$

Вопрос 13

В 2 раза.

Вопрос 14

Предположим, что ошибки наблюдения распределены по нормальному закону

$N(0, \sigma^2 I_N)$. Пусть априорные данные о параметрах модели определены информацией из выборки A и состоят в выполнении $X_A^T X_A \theta = X_A^T y_A$. Поступают новые данные в виде выборки B . По ней производится оценивание параметров и осуществляется проверка гипотезы H : $X_A^T X_A \theta = X_A^T y_A$.

Обозначим

$$RSS = \|y_B - \hat{y}_B(B)\|^2 = (y_B - X_B \hat{\theta}_B)^T (y_B - X_B \hat{\theta}_B) \equiv \varepsilon^2(B), \quad (1)$$

$$RSS_H = (y_B - X_B \hat{\theta}_H)^T (y_B - X_B \hat{\theta}_H), \quad (2)$$

где $\hat{\theta}_H$ - оценки, полученные при ограничениях $X_A^T X_A \theta = X_A^T y_A$.

В нашем случае при условии полного ранга матрицы $X_A^T X_A$ имеем

$$\hat{\theta}_H = \hat{\theta}_A = (X_A^T X_A)^{-1} X_A^T y_A.$$

Статистика F для проверки гипотезы H будет следующая

$$F = \frac{(RSS_H - RSS) / p}{RSS / (N_B - p)},$$

где p – число параметров в модели. С учетом ранее введенных обозначений F -статистику можно переписать

$$F = \frac{(\Delta_{reg}^2(B) - \varepsilon^2(B)) / p}{\varepsilon^2(B) / (N_B - p)} = \frac{\Delta_{reg}^2(B)}{\varepsilon^2(B)} - \frac{N_B - p}{p}.$$

Вопрос 15

Под состоятельностью понимается сходимость по вероятности

$$\lim_{N \rightarrow \infty} P \left[(\hat{\theta}_N - \theta)^T (\hat{\theta}_N - \theta) \geq \gamma \right] = 0.$$

Состоятельность оценки (точнее – последовательности оценок) означает концентрацию вероятностной массы около истинного значения параметра с ростом объема выборки. В многомерном случае требуется, чтобы состоятельность выполнялась для каждого компонента θ_i вектора θ .

Будем рассматривать состоятельность в среднем квадратичном:

$$E(\hat{\theta}_N - \theta)^T (\hat{\theta}_N - \theta) \rightarrow 0, N \rightarrow \infty.$$

Сходимость в среднеквадратичном влечет сходимость по вероятности. Это сразу вытекает из неравенства Чебышева

$$P \left[|\hat{\theta}_N - \theta| \geq \gamma \right] \leq \frac{1}{\gamma} E(\hat{\theta}_N - \theta)^2.$$

Величина $E(\hat{\theta}_N - \theta)^T (\hat{\theta}_N - \theta)$ есть след матрицы ковариаций $E(\hat{\theta}_N - \theta)(\hat{\theta}_N - \theta)^T \equiv D(\hat{\theta}_N)$.

Таким образом для выполнения условия

$tr D(\hat{\theta}_N) \equiv E(\hat{\theta}_N - \theta)^T (\hat{\theta}_N - \theta) \rightarrow 0, N \rightarrow \infty$ необходимо потребовать выполнение условия

$$D(\hat{\theta}_N) \rightarrow O, N \rightarrow \infty.$$

Поскольку $D(\hat{\theta}_N) = \sigma^2 \left(\sum_{i=1}^N f(x_i) f^T(x_i) \right)^{-1} = \sigma^2 (X^T X)^{-1}$ необходимо чтобы матрица $X^T X$ бесконечно возрастала с ростом N . Поскольку $f(x_i) f^T(x_i)$ - п.п.о., то всегда найдется такое k , что $\sigma^2 \sum_{i=1}^N f(x_i) f^T(x_i) < \sigma^2 \sum_{i=1}^{N+k} f(x_i) f^T(x_i)$, что означает неограниченный рост $X^T X$ ■

Задача

1. Модель $\eta_1 = \theta_1 + \theta_2 x^2$
2. Гипотеза принимается

Билет №22

Вопрос 1

При активном эксперименте исследователь сам устанавливает необходимые комбинации значений входных факторов.

Вопрос 2

Адекватность модели – способность описывать поведение объекта (системы) в среднем в условиях того, что экспериментальные наблюдения за откликом подвержены влиянию случайных помех.

Вопрос 3

Доверительное оценивание для отдельного параметра

Доверительный интервал для отдельного параметра также можно получить исходя из общего выражения:

$$(\Theta_j - \hat{\Theta}_j)^T \frac{1}{d_{jj}} (\Theta_j - \hat{\Theta}_j) \leq \hat{\sigma}^2 F_{\alpha, 1, n-m}$$

Учитывая, что $\left((X^T X)^{-1} \right)_{jj} \sigma^2 = \sigma^2 \left(\hat{\Theta}_j \right)$ - дисперсия оценки отдельного параметра, и что ,

где $t_{\alpha/2, f_R}$ - t -критерий Стьюдента, $f_R = n-m$, то можно записать в виде: $\frac{|\Theta_j - \hat{\Theta}_j|}{\sigma(\hat{\Theta}_j)} \leq t_{\alpha/2, f_R}$.

Или в виде двухстороннего неравенства:

$$\hat{\Theta}_j - t_{\alpha/2, f_R} \sigma(\hat{\Theta}_j) \leq \Theta_j \leq \hat{\Theta}_j + t_{\alpha/2, f_R} \sigma(\hat{\Theta}_j).$$

Вопрос 4

Выделяющиеся точки графика: некоторые из остатков могут по абсолютной величине сильно превосходить все остальные остатки. Если максимальное значение остатка $|d_{\max}| > \delta$, где δ - некоторое заранее выбранное число, то наблюдение, имеющее такой остаток, является аномальным. Величина остатка аномального наблюдения зависит от объема выборки n : при $n > 20$ его можно обнаружить по правилу «трех сигм» ($\delta = 3$), при $n < 20$, например, по критерию Смирнова - Граббса. В этом случае вычисляемая статистика имеет вид

$$D = (d_{\max} - \bar{d}) / S_d, \text{ где } \bar{d} - \text{среднее значение остатков, } S_d^2 = 1/n \left(\sum_{i=1}^n (d_i - \bar{d})^2 \right).$$

Вопрос 5

Ридж-оценка имеет минимальную длину в классе оценок с заданным значением суммы квадратов отклонений (RSS).

Вопрос 6

Статистика Мэллоуса, коэффициент детерминации, критерий Акаике.

Вопрос 7

Для критерия Δ_{abs}^2 , например, его среднее значение будет определяться величиной

$$\bar{\Delta}_{abs}^2 = u^T D_{A_{abs}(p)} u + \sigma^2 \cdot p.$$

С увеличением σ^2 минимум критерия в его зависимости от сложности модели смещается влево, т.е. выбираются все более простые модели. Абсолютно помехоустойчивый критерий получил свое название по причине того, что второе слагаемое для него менее всего возрастает при увеличении дисперсии помехи. Другими словами, при увеличении дисперсии помехи минимум данного критерия с наименьшей скоростью смещается влево в сторону все более простых моделей.

Вопрос 8

На первом шаге выбирается некоторое число лучших моделей, состоящих из одного регрессора. На втором этапе к этим моделям добавляются различные регрессоры по одному, и отбираются модели с двумя аргументами, улучшающие величину критерия, и т.д. до его ухудшения или до построения полной модели.

На последнем шаге можно из оставшихся лучших моделей выбрать одну по другому критерию.

Вопрос 9

Для линейной модели с двумя факторами при 4 наблюдениях в углах квадрата информационная матрица будет диагональной с диагональю 4. При 8 наблюдениях (повторы) диагональ будет равна 8 и т.д.

Вопрос 10

Имеем $X^T(y - X\hat{\theta}) = 0$ и далее $X^T(y - \hat{y}) = 0$, т.е. вектор остатков $\hat{e} = y - \hat{y}$ ортогонален Ω . В тоже время справедливо представление $y = \hat{y} + \hat{e}$. Становится понятным, что $\hat{y}$ есть проекция y на $R(X)$, а именно $\hat{y} = Py$. Можно рассматривать также подпространство $R^\perp(X)$ – ортогональное дополнение для $R(X)$ в $R(X)$. Поскольку

$$\hat{e} = y - \hat{y} = y - X\hat{\theta} = (I_n - X(X^T X)^{-1} X^T) y = (I_n - P) y,$$

то $M = I_n - P$ – есть матрица проектирования вектора y на подпространство $R^\perp(X)$.

Вопрос 11

$$\hat{\Theta} = (X^T X)^{-1} X^T y. \quad \hat{\Theta} = (n)^{-1} X^T y = \sum_{i=1}^n y_i / n$$

Вопрос 12

Для наглядности рассмотрим случай $k=2$.

$$y = \theta_1 x_1 + \theta_2 x_2 + \varepsilon.$$

Пусть для удобства мы пронормировали вектора

$$\underline{x}_1^T = (x_{11}, \dots, x_{1n}), \quad \underline{x}_2^T = (x_{21}, \dots, x_{2n}): \quad \underline{x}_1' = \frac{\underline{x}_1}{\|\underline{x}_1\|}, \quad \underline{x}_2' = \frac{\underline{x}_2}{\|\underline{x}_2\|}.$$

Тогда

$$X^T X = \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}, \text{ где } r_{12} = \cos(\underline{x}_1, \underline{x}_2) = \sum \frac{x_{1t} x_{2t}}{\sqrt{\sum x_{1t}^2 \cdot \sum x_{2t}^2}}.$$

Тогда

$$(X^T X)^{-1} = \begin{pmatrix} \frac{1}{(1-r_{12}^2)} & -\frac{r_{12}}{(1-r_{12}^2)} \\ -\frac{r_{12}}{(1-r_{12}^2)} & \frac{1}{(1-r_{12}^2)} \end{pmatrix} \text{ и } \sigma^2(\hat{\theta}_1) = \sigma^2(\hat{\theta}_2) = \frac{\sigma^2}{(1-r_{12}^2)}.$$

Если между переменными x_1 и x_2 существует некая линейная зависимость, т.е. r_{12}^2 близко к 1, то, как следует из последнего выражения, дисперсии оценок параметров будут иметь большие значения. При $r_{12} \rightarrow 1 \Rightarrow \sigma^2(\hat{\theta}_1) \rightarrow \infty$.

Вопрос 13

- 2) $\hat{\beta}(\lambda)$ - является линейным преобразованием МНК-оценки и является смещённой;
Действительно,

$$Z^T y = (I + \lambda(Z^T Z)^{-1})^{-1} (Z^T Z)^{-1} Z^T y = B \hat{\beta},$$

где $B = (I + \lambda(Z^T Z)^{-1})^{-1}$ - матрица линейного преобразования и $B \neq I$, что приводит к смещению оценок $\hat{\beta}(\lambda)$ в отличие от $\hat{\beta}$, которые несмещены.

Вопрос 14

При переусложнении модели имеется дополнительная возможность описывать не только среднее поведение системы, но и флуктуации, обусловленные влиянием помех. На новых тестовых данных помеха уже будет другая, что и создаст условия увеличения ошибки предсказания.

Вопрос 15

Введем в рассмотрение вектор остатков: -

$$\hat{e} = y - \hat{y} = y - X \hat{\theta} = (I_n - X(X^T X)^{-1} X^T) y = (I_n - P) y,$$

где P - матрица проектирования: $\hat{y} = P y$, т.е. y проектируется на Ω ; $(I_n - P)$ - матрица ортогонального проектирования на $\Omega^\perp$, где $\Omega = R[X]$ - пространство образов, порождаемое столбцами матрицы X .

Матрицы P и $I_n - P$ симметричны и идемпотентны, т.е. $P^2 = P$, $(I_n - P)^2 = (I_n - P)$.

Поскольку матрица $I_n - P$ симметрична и идемпотентна, то $\text{rang}[I_n - P] = \text{tr}[I_n - P] = n - \text{tr}[P]$ т.е. у идемпотентной матрицы ранга r имеется r собственных чисел равных 1, а остальные $n-r$ равны 0. В нашем случае:

$$\text{tr}[P] = \text{tr}[X(X^T X)^{-1} X^T] = \text{tr}[X^T X (X^T X)^{-1}] = \text{tr}[I_m] = m.$$

Введем в рассмотрение статистику

$$\hat{e}^T \hat{e} = (y - X\hat{\theta})^T (y - X\hat{\theta}) = y^T (I_n - P)^T (I_n - P) y = y^T (I_n - P)^2 y.$$

Теорема о математическом ожидании квадратичной формы от случайной величины:

Пусть Z - случайный вектор размера $n \times 1$, A - симметричная $n \times n$ матрица. Если $E(Z) = \xi$ и $D(Z) = \Sigma$, то $E[Z^T A Z] = \text{tr}[A \Sigma] + \xi^T A \xi$.

Применительно к нашему случаю имеем:

$$E[\hat{e}^T \hat{e}] = E[y^T (I_n - P) y] = \sigma^2 \text{tr}(I_n - P) + \theta_{\hat{\theta}}^T X^T (I_n - P) X \theta_{\hat{\theta}} = \sigma^2 (n - m).$$

Примечание. Второе слагаемое здесь равно нулю поскольку $(I_n - P)X = 0$.

Отсюда заключаем, что статистика $\hat{\sigma}^2 = \frac{\hat{e}^T \hat{e}}{n - m}$ - несмещенная оценка для σ^2 .

Задача

1. Модель $\eta_1 = \theta_1 + \theta_2 x^2$
2. Гипотеза отвергается

Билет №23

Вопрос 1

НЛО, МНК, ММП, Ридж-оценивание, ОМНК

Вопрос 2

Для возможности обращения информационной матрицы

Вопрос 3

Для ее вычисления необходимо обратить матрицу коэффициентов корреляций, которая может быть плохо обусловлена в условиях мультиколлинеарности.

Вопрос 4

Ридж-оценки впервые были введены А. Гоэрлом при исследовании регрессии 2-го порядка (ridge – гребень, хребет). Название "ридж-оценки" навеяно следующими наблюдениями. При сильной мультиколлинеарности линии равного уровня доверительного эллипсоида для разного уровня значимости (доверительной вероятности) становятся сильно вытянутыми и визуально похожими на хребет на географической карте. Данные оценки использовались именно в условиях мультиколлинеарности и поэтому получили название ридж-оценок.

Вопрос 5

Ошибка спецификации модели – это недобор или перебор регрессоров.

Вопрос 6

Критерии точности.

Типичным критерием этого вида является критерий регулярности:

$$\begin{aligned}\Delta_{reg}^2(B) &= \Delta^2(B/A) = \|y_B - \hat{y}_B(A)\|^2 = \\ &= (y_B - X_B \hat{\theta}_A)^T (y_B - X_B \hat{\theta}_A) = \|y_B - X_B \hat{\theta}_A\|^2\end{aligned}$$

где запись $\Delta^2(B/A)$ означает «ошибка на выборке B модели, коэффициенты которой получены на A », $\hat{\theta}_A = (X_A^T X_A)^{-1} X_A^T y_A$, $\hat{y}_B(A) = X_B \hat{\theta}_A$. Этот критерий несимметричен.

Вопрос 7

Случай недобора.

Пусть истинное уравнение (4.2), а мы оцениваем регрессию

$$y_u = \gamma_1 \omega_{u1} + \dots + \gamma_k \omega_{uk} + \xi_u \quad (4.5)$$

или

$$y = W\gamma + \xi, \quad X = \begin{bmatrix} W & V \end{bmatrix}$$

Другими словами, $\omega_1 = x_1, \dots, \omega_k = x_k$ входят в регрессию (4.5), а остальные $m-k$ переменных $\psi = x_{k+1}, \dots, \psi_{m-k} = x_m$ в регрессии отсутствуют.

Докажем, что тогда используемая оценка

$$\hat{\gamma} = (W^T W)^{-1} W^T y \quad (4.6)$$

в общем случае является смещённой в том смысле, что $E\hat{\gamma} \neq \alpha_1$.

Действительно,

$$E\hat{\gamma} = (W^T W)^{-1} W^T E y = (W^T W)^{-1} W^T X \alpha = (W^T W)^{-1} W^T \begin{bmatrix} W & V \end{bmatrix} \cdot \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix},$$

где α_1 и α_2 - подвекторы вектора α размерности k и $m-k$ соответственно.

Перемножая, получим

$$E\hat{\gamma} = \begin{bmatrix} I_k & (W^T W)^{-1} W^T V \end{bmatrix} \cdot \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \alpha_1 + (W^T W)^{-1} W^T V \alpha_2 \quad \blacksquare$$

Оценка $\hat{\gamma}$ будет несмещённой, когда матрицы W и V ортогональны, то есть $W^T V = 0$.

Можно показать также, что МНК оценка в случае недобора является несостоятельной.

Вопрос 8

Для доказательства несмещенности оценок.

Вопрос 9

По причинам неадекватности модели, обнаружения мультиколлинеарности.

Вопрос 10

$$P = X(X^T X)^{-1} X^T, \quad P^2 = X(X^T X)^{-1} X^T X(X^T X)^{-1} X^T = P$$

Вопрос 11

Точечную оценку дисперсии ошибки эксперимента в случае, когда имеется повторные опыты лучше всего определять через сумму (2.23) по формуле

$$\hat{\sigma}_E^2 = SS_E / f_E, \text{ где } f_E \text{ число степеней свободы суммы } SS_E, \text{ равное}$$

$f_E = N - N_c = \sum_{v=1}^{N_c} (r_v - 1)$. Оценка $\hat{\sigma}_E^2$ есть не что иное, как объединенная оценка

выборочных дисперсий ошибки опыта, которые можно найти по результатам отдельных серий повторных опытов. Оценка дисперсии ошибки в v -ой серии повторных опытов равна

$$\hat{\sigma}_{Ev}^2 = \frac{SS_{Ev}}{f_{Ev}}, \text{ где } SS_{Ev} = \sum_{j=1}^{r_v} (y_{vj} - \bar{y}_v)^2, f_{Ev} = r_v - 1.$$

Отсюда следует, что объединенная оценка по всем сериям имеет вид

$$\hat{\sigma}_E^2 = \sum_{v=1}^{N_c} \frac{f_{Ev}}{f_E} \hat{\sigma}_{Ev}^2 = \sum_{v=1}^{N_c} SS_{Ev} / f_E = SS_E / f_E.$$

Вопрос 12

. Рассмотрим случай когда в модели присутствует свободный член. В этом случае сумма остатков равно нулю, т.е.

$$\sum (y_i - \hat{y}_i) = 0. \quad (a)$$

Покажем это. Для этого сумму квадратов отклонений для модели со свободным членом

$$\sum (y_i - \theta_0 - \theta_1 x_{1i} - \dots - \theta_{m-1} x_{m-1,i})^2$$

продифференцируем по θ_0 :

$$\sum (y_i - \hat{\theta}_0 - \hat{\theta}_1 x_{1i} - \dots - \hat{\theta}_{m-1} x_{m-1,i}) = 0 \text{ или } \sum (y_i - \hat{y}_i) = 0.$$

Вопрос 13

Ясно, что в реальных ситуациях редко бывает так, что параметр автокорреляции ρ известен. При неизвестном ρ процедуры оценивания параметров θ носят итеративный характер.

Процедура Кохрейна-Оркатта.

6. Для исходной модели оцениваются $\hat{\theta}$ и $\hat{e}$;
7. Рассматривается регрессия $\hat{e}_t = \rho \hat{e}_{t-1} + v_t$ и вычисляется оценка $\hat{\rho}$;
8. Проводится ОМНК оценивание исходной модели с $\hat{\Omega}^{-1}$, где в качестве ρ используется его оценка $\hat{\rho}$;
9. Вычисляется новый вектор остатков $\hat{e}$;
10. Процедура повторяется с п.2.

Останов производится после нескольких итераций или после того как очередное $\hat{\rho}$ мало отличается от предыдущего.

Процедура Хилдрета-Лу.

По схеме ОМНК производится решение следующей экстремальной задачи:

$$\hat{\rho} = \text{Arg} \min_{\rho \in [-1, +1]} RSS.$$

Попутно при этом мы получаем оценки для вектора параметров θ .

Вопрос 14

$$p_{ii} = f^T(x_i)(X^T X)^{-1} f(x_i)$$

. На практике руководствуются следующими правилами. Значения $p_{ii} \leq 0,2$ позволяют считать надежными данные наблюдения, значения $0,2 \leq p_{ii} \leq 0,5$ относят данное наблюдение к классу рискованных, а при значениях $p_{ii} > 0,5$ i -е наблюдение можно принять за точку разбалансировки.

Вопрос 15

Пусть как и ранее модель наблюдения имеет вид:

$$y = X\Theta + \varepsilon,$$

где X - матрица $n \times m$ ранга m .

Предположим, что мы хотим найти оценки параметра Θ , минимизируя $\varepsilon^T \varepsilon$, при наличии линейных ограничений $A\Theta = c$, где A - известная $q \times m$ матрица ранга $q < m$, а c - известный вектор $q \times 1$.

Применим метод множителей Лагранжа. Рассмотрим целевую функцию:

$$r = \varepsilon^T \varepsilon + (\Theta^T A^T - c^T) \lambda,$$

где $\lambda^T = (\lambda_1, \dots, \lambda_q)$ - вектор множителей Лагранжа. Решим уравнения:

$$A\Theta = c \tag{2.1}$$

$$\frac{\partial r}{\partial \Theta} = -2X^T y + 2X^T X\Theta + A^T \lambda = 0. \tag{2.2}$$

Решения уравнений (2.1) и (2.2) обозначим через $\hat{\lambda}_H$ и $\hat{\Theta}_H$. Тогда из (2.2) получаем:

$$\hat{\Theta}_H = (X^T X)^{-1} X^T y - \frac{1}{2} (X^T X)^{-1} A^T \hat{\lambda}_H = \hat{\Theta} - \frac{1}{2} (X^T X)^{-1} A^T \lambda_H, \tag{2.3}$$

а из (2.1):

$$c = A\hat{\Theta}_H = A\hat{\Theta} - \frac{1}{2} A(X^T X)^{-1} A^T \hat{\lambda}_H.$$

Поскольку матрица $(X^T X)^{-1}$ - п.п.о., то матрица $A(X^T X)^{-1} A^T$ будет п.п.о. ранга q . Следовательно

$$-\frac{1}{2} \hat{\lambda}_H = [A(X^T X)^{-1} A^T]^{-1} (c - A\hat{\Theta}).$$

Окончательно подставляя в (2.3):

$$\hat{\Theta}_H = \hat{\Theta} + (X^T X)^{-1} A^T [A(X^T X)^{-1} A^T]^{-1} (c - A\hat{\Theta}). \tag{2.4}$$

Задача

1. Модель $\eta_1 = \theta_1 x + \theta_2 x^2$
2. Гипотеза отвергается

Билет №24

Вопрос 1

Структура модели – это набор регрессоров.

Вопрос 2

Оценки параметров распределены по нормальному закону с мат.ожиданием равным истинным значениям параметров и дисперсией $\sigma^2(X^T X)^{-1}$

Вопрос 3

Это проверка того, что модель системы не изменилась со временем.

Вопрос 4

2) $\hat{\beta}(\lambda)$ - является линейным преобразованием МНК-оценки и является смещённой;
Действительно,

$$Z^T y = (I + \lambda(Z^T Z)^{-1})^{-1} (Z^T Z)^{-1} Z^T y = B \hat{\beta},$$

где $B = (I + \lambda(Z^T Z)^{-1})^{-1}$ - матрица линейного преобразования и $B \neq I$, что приводит к смещению оценок $\hat{\beta}(\lambda)$ в отличие от $\hat{\beta}$, которые несмещены.

Вопрос 5

Под помехоустойчивостью внешнего критерия будем понимать его способность выбирать из всего множества моделей, построенных по зашумленным данным, такую модель оптимальной сложности, которая достаточно точно восстанавливает не зашумленный выход объекта.

Вопрос 6

$$I - P = I - X(X^T X)^{-1} X^T.$$

$$(I - P)^2 = I - 2P + P^2 = \\ = I - 2X(X^T X)^{-1} X^T + X(X^T X)^{-1} X^T X(X^T X)^{-1} X^T = I - P$$

Вопрос 7

Частный случай 1: проверка незначимости отдельного параметра $H: \theta_j = c$ или $H:$

$a^T \theta = c$, где a^T - вектор-строка, в которой на j -ом месте стоит 1, на остальных местах – нули. Обозначим $d_{jj} = ((X^T X)^{-1})_{jj}$ - j -ый диагональный элемент. Тогда

$$F = \frac{(a^T \hat{\theta} - c)^T (a^T \hat{\theta} - c)}{\hat{\sigma}^2 d_{jj}} = \frac{(\hat{\theta}_j - c)^2}{\hat{\sigma}^2 d_{jj}}$$

имеет при гипотезе H распределение $F_{1, n-m}$. т.е если $F \leq F_{\alpha, 1, n-m}$, где $F_{\alpha, 1, n-m}$ - критическая точка, то гипотеза о незначимости не отклоняется.

Вопрос 8

Доступный обобщенный МНК. В случае обнаружения эффекта гетероскедастичности можно провести коррекцию оценок параметров модели с помощью взвешенного МНК с весами $\hat{\sigma}_t^{-2} = h^{-1}(z_t^T \hat{\alpha})$. Может оказаться, что для некоторых t $h^{-1}(z_t^T \hat{\alpha}) < 0$. Если число таких наблюдений невелико, то их можно просто выбросить, в противном случае использовать представление $\sigma_t^2 = \exp(h(z_t^T \alpha))$.

Вопрос 9

В качестве критерия величины смещения (функции риска) выберем среднюю сумму квадратов ошибок (ССКО)

$$L_{\tilde{\theta}}(\theta) = E(\tilde{\theta} - \theta)^T (\tilde{\theta} - \theta),$$

где $\tilde{\theta}$ - некоторая оценка параметров $\theta \in R^m$.

Для МНК оценки

$$L_{\hat{\theta}}(\theta) = \sigma^2 \text{tr}(X^T X)^{-1} \rightarrow \infty \text{ при } \lambda_1 \rightarrow 0,$$

т.е. МНК оценка в условиях усиления коллинеарности ($\lambda_1 \rightarrow 0$) не обеспечивает ограниченность ССКО. Это будет происходить и в условиях увеличения числа наблюдений. В нормальных условиях (при отсутствии мультиколлинеарности) $\text{tr}(X^T X)^{-1} \rightarrow 0$ при $N \rightarrow \infty$, т.е. теоретически гарантируется состоятельность в среднеквадратичном.

Вопрос 10

Пусть истинное уравнение (4.2), а мы оцениваем регрессию

$$y_u = \gamma_1 \omega_{u1} + \dots + \gamma_k \omega_{uk} + \xi_u \quad (4.5)$$

или

$$y = W\gamma + \xi, \quad X = \begin{bmatrix} W & V \end{bmatrix}$$

Другими словами, $\omega_1 = x_1, \dots, \omega_k = x_k$ входят в регрессию (4.5), а остальные $m - k$ переменных $\psi = x_{k+1}, \dots, \psi_{m-k} = x_m$ в регрессии отсутствуют.

Докажем, что тогда используемая оценка

$$\hat{\gamma} = (W^T W)^{-1} W^T y \quad (4.6)$$

в общем случае является смещённой в том смысле, что $E\hat{\gamma} \neq \alpha_1$.

Действительно,

$$E\hat{\gamma} = (W^T W)^{-1} W^T E y = (W^T W)^{-1} W^T X \alpha = (W^T W)^{-1} W^T \begin{bmatrix} W & V \end{bmatrix} \cdot \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix},$$

где α_1 и α_2 - подвекторы вектора α размерности k и $m - k$ соответственно.

Перемножая, получим

$$E\hat{\gamma} = \begin{bmatrix} I_k & (W^T W)^{-1} W^T V \end{bmatrix} \cdot \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \alpha_1 + (W^T W)^{-1} W^T V \alpha_2 \quad \blacksquare$$

Оценка $\hat{y}$ будет несмещённой, когда матрицы W и V ортогональны, то есть $W^T V = 0$.

Можно показать также, что МНК оценка в случае недобора является несостоятельной.

Вопрос 11

В таблице исходных данных могут присутствовать наблюдения (например, наблюдение под номером i), сильно влияющие на оценки параметров. Если эти наблюдения связаны с матрицей X , то наблюдение x_i называется влияющим или точкой разбалансировки. Точки разбалансировки не обязательно дают большие остатки в соответствующих наблюдениях, эти большие остатки могут появиться в других наблюдениях.

Вопрос 12

В 4 раза.

Вопрос 13

Перебор регрессоров приводит к неэффективности оценок. Недобор – к смещению.

Вопрос 14

Критерий непротиворечивости, абсолютно помехоустойчивый критерий.

Вопрос 15

Рассмотрим более общий случай, когда среди опытов, входящих в эксперимент, имеются повторные.

Без ограничения общности можно сгруппировать опыты по сериям так чтобы каждая из них содержала все повторные опыты, которые проведены при одних и тех же условиях.

Положим, что всего имеются N_c серий и в каждую v -ю серию входят r_v - повторных опытов,

$v = \overline{1, N_c}$. Общее число опытов $N = \sum_{v=1}^{N_c} r_v$

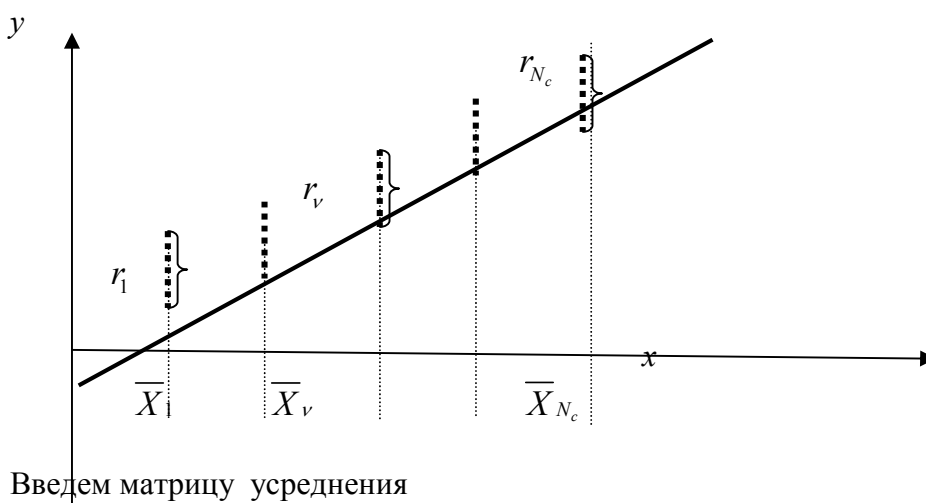
Как и раньше рассматриваем линейную модель наблюдения

$$y(x) = f^T(x)\theta + \varepsilon \quad (2.14)$$

или в матричной форме

$$y = X\theta + \varepsilon. \quad (2.15)$$

Будем считать, что элементы векторов y, ε и строки матрицы X сгруппированы по сериям.



Введем матрицу усреднения

$$S = \begin{bmatrix} r_1^{-1} I_{r_1}^T & & & 0 \\ & r_2^{-1} I_{r_2}^T & & \\ & & \ddots & \\ 0 & & & r_{N_c}^{-1} I_{r_{N_c}}^T \end{bmatrix} \text{ размером } N_c \times N,$$

где I_{r_v} - вектор столбец размерности r_v , состоящий из единиц. Из структуры этой матрицы видно, что

$$SS^T = R^{-1} = (r_i^{-1} \delta_{ij}), \quad i, j = \overline{1, N_c}, \quad (2.16)$$

где R - диагональная матрица весов, выражающая распределения опытов по сериям.

Умножим теперь левую и правую часть (2.15) слева на матрицу S и получим

$$\bar{y} = \bar{X}\theta + \varepsilon. \quad (2.17)$$

Легко понять, что вектор $\bar{y}$ в качестве элементов имеет среднее значения отклика по сериям:

$$\bar{y} = Sy = (\bar{y}_1, \bar{y}_2, \dots, \bar{y}_{N_c})^T, \text{ причем } \bar{y}_v = \bar{y}(x_v) = r_v^{-1} \sum_{j=1}^{r_v} y_{v_j}, v = \overline{1, N_c}.$$

Матрица $\bar{X}$ содержит N_c строк, которые отличаются друг от друга и соответствуют разным сериям:

$$\bar{X} = SX = (f(x_1), f(x_2), \dots, f(x_{N_c}))^T.$$

Вектор усреднённых ошибок имеет вид $\bar{\varepsilon} = (\bar{\varepsilon}_1, \dots, \bar{\varepsilon}_{N_c})^T$, где $\bar{\varepsilon}_v = r_v^{-1} \sum_{j=1}^{r_v} \varepsilon_{v_j}$, $v = \overline{1, N_c}$.

Учитывая, что

$$E(\varepsilon) = O_{N,1}; \quad E(\varepsilon\varepsilon^T) = D(\varepsilon) = I_N \sigma^2,$$

Получим

$$E(\bar{\varepsilon}) = O_{N_c,1}; \quad E(\bar{\varepsilon}\bar{\varepsilon}^T) = D(\bar{\varepsilon}) = R^{-1} \sigma^2 = SI_N \sigma^2 S^T.$$

Сумму взвешенных средних квадратичных отклонений для модели (2.15) можно записать в виде квадратичной формы

$$S(\theta) = \varepsilon^T D^{-1}(\varepsilon) \varepsilon = (y - X\theta)^T \sigma^{-2} (y - X\theta).$$

Опуская сомножитель σ^{-2} , запишем

$$SS(\theta) = (y - X\theta)^T (y - X\theta). \quad (2.18)$$

Для модели (2.17) аналогичная мера рассогласования будет иметь вид

$$S_1(\theta) = \bar{\varepsilon}^T D^{-1}(\bar{\varepsilon}) \bar{\varepsilon} = (\bar{y} - \bar{X}\theta)^T R \sigma^{-2} (\bar{y} - \bar{X}\theta).$$

Опуская σ^{-2} , запишем

$$SS_1(\theta) = (\bar{y} - \bar{X}\theta)^T R (\bar{y} - \bar{X}\theta). \quad (2.19)$$

Матрица весов R может быть представлена в виде произведения

$$R = Q^T Q, \quad (2.20)$$

где Q - матрица размером $N \times N_c$

$$Q = \begin{bmatrix} I_{r_1} & . & . & . & 0 \\ . & I_{r_2} & & & . \\ . & & . & & . \\ . & & & . & . \\ 0 & . & . & . & I_{r_{N_c}} \end{bmatrix}, \text{ где } I_{r_v} - \text{ столбец единиц}$$

Подставляя (2.20) в (2.19) получим

$$SS_1(\theta) = (Q\bar{y} - X\theta)^T (Q\bar{y} - X\theta), \quad (2.21)$$

где вектор $Q\bar{y}$ – имеет уже N – составляющих, как и вектор y .

Суммы $SS(\theta)$ и $SS_1(\theta)$ связаны между собой. Используя перегруппировку слагаемых первой суммы, можно показать, что

$$SS(\theta) = SS_1(\theta) + SS_E, \quad (2.22)$$

где

$$SS_E = (y - Q\bar{y})^T (y - Q\bar{y}) \quad (2.23)$$

и не зависима от θ . Поэтому совершенно безразлично какой мерой рассогласования пользоваться для отыскания оценок параметров: (2.18), (2.19) или (2.21).

Точечную оценку дисперсии ошибки эксперимента в случае, когда имеется повторные опыты лучше всего определять через сумму (2.23) по формуле

$\hat{\sigma}_E^2 = SS_E / f_E$, где f_E число степеней свободы суммы SS_E , равное

$f_E = N - N_c = \sum_{v=1}^{N_c} (r_v - 1)$. Оценка $\hat{\sigma}_E^2$ есть не что иное, как объединенная оценка

выборочных дисперсий ошибки опыта, которые можно найти по результатам отдельных серий повторных опытов. Оценка дисперсии ошибки в v -ой серии повторных опытов равна

$$\hat{\sigma}_{Ev}^2 = \frac{SS_{Ev}}{f_{Ev}}, \text{ где } SS_{Ev} = \sum_{j=1}^{r_v} (y_{vj} - \bar{y}_v)^2, \quad f_{Ev} = r_v - 1.$$

Отсюда следует, что объединенная оценка по всем сериям имеет вид

$$\hat{\sigma}_E^2 = \sum_{v=1}^{N_c} \frac{f_{Ev}}{f_E} \hat{\sigma}_{Ev}^2 = \sum_{v=1}^{N_c} SS_{Ev} / f_E = SS_E / f_E.$$

Заметим, что оценка $\hat{\sigma}_E^2$ определяется независимо от модели.

При $\theta = \hat{\theta}$ имеем следующие соотношения:

$$SS_R \equiv SS(\hat{\theta}) = SS_1(\hat{\theta}) + SS_E, \quad (2.24)$$

где SS_R - остаточная сумма со степенями свободы $f_R = N - m$;

$SS_{LF} = SS_1(\hat{\theta})$ - мера степени неадекватности представления экспери

ментальных данных с помощью выбранной модели. Она оценивает рассеивания средних

значений отклика по сериям относительно расчетной регрессионной зависимости с учётом весов (**Lack of fit – неадекватность**).

Оценка дисперсии на основе суммы SS_{LF} вычисляется по формуле

$$\hat{\sigma}_{LF}^2 = SS_{LF} / f_{LF}, \quad f_{LF} = N_c - m. \quad (2.25)$$

Оценка величины σ^2 на основе SS_R выражается отношением

$$\hat{\sigma}_R^2 = SS_R / f_R, \quad f_R = N - m.$$

Оценка величины $\hat{\sigma}_R^2$ является, по существу, объединённой. Используя (2.24) и учитывая, что $f_R = f_{LF} + f_E$ можно записать

$$\hat{\sigma}_R^2 = (f_{LF} / f_R) \hat{\sigma}_{LF}^2 + (f_E / f_R) \hat{\sigma}_E^2.$$

Задача

1. Модель $\eta_1 = \theta_1 x + \theta_2 x^2$
2. Гипотеза принимается

Билет №25

Вопрос 1

Часто класс допустимых решений задают в виде параметрического множества функций.

Вопрос 2

Оценки параметров распределены по нормальному закону с мат.ожиданием равным истинным значениям параметров и дисперсией $\sigma^2 (X^T X)^{-1}$

Вопрос 3

$$D(\hat{\theta}) = E(\hat{\theta} - E\hat{\theta})(\hat{\theta} - E\hat{\theta})^T = E[(X^T X)^{-1} X^T \varepsilon \varepsilon^T X (X^T X)^{-1}] = \sigma^2 (X^T X)^{-1} \blacksquare$$

Вопрос 4

Это проверка того, что модель системы не изменилась со временем.

Вопрос 5

Число главных компонент определяется по числу относительно больших собственных значений информационной матрицы

Вопрос 6

$$I - P = I - X(X^T X)^{-1} X^T.$$

$$(I - P)^2 = I - 2P + P^2 =$$

$$= I - 2X(X^T X)^{-1} X^T + X(X^T X)^{-1} X^T X (X^T X)^{-1} X^T = I - P$$

Вопрос 7

Для линейной модели с двумя факторами при 4 наблюдениях в углах квадрата информационная матрица будет диагональной с диагональю 4. При 8 наблюдениях (повторы) диагональ будет равна 8 и т.д.

Вопрос 8

Доступный обобщенный МНК. В случае обнаружения эффекта гетероскедастичности можно провести коррекцию оценок параметров модели с помощью взвешенного МНК с весами $\hat{\sigma}_i^{-2} = h^{-1}(z_i^T \hat{\alpha})$. Может оказаться, что для некоторых i $h^{-1}(z_i^T \hat{\alpha}) < 0$. Если число таких наблюдений невелико, то их можно просто выбросить, в противном случае использовать представление $\sigma_i^2 = \exp(h(z_i^T \alpha))$.

Вопрос 9

В качестве критерия величины смещения (функции риска) выберем среднюю сумму квадратов ошибок (ССКО)

$$L_{\tilde{\theta}}(\theta) = E(\tilde{\theta} - \theta)^T (\tilde{\theta} - \theta),$$

где $\tilde{\theta}$ - некоторая оценка параметров $\theta \in R^m$.

Для МНК оценки

$$L_{\hat{\theta}}(\theta) = \sigma^2 \text{tr}(X^T X)^{-1} \rightarrow \infty \text{ при } \lambda_1 \rightarrow 0,$$

т.е. МНК оценка в условиях усиления коллинеарности ($\lambda_1 \rightarrow 0$) не обеспечивает ограниченность ССКО. Это будет происходить и в условиях увеличения числа наблюдений. В нормальных условиях (при отсутствии мультиколлинеарности) $\text{tr}(X^T X)^{-1} \rightarrow 0$ при $N \rightarrow \infty$, т.е. теоретически гарантируется состоятельность в среднеквадратичном.

Вопрос 10

Переменные, которым соответствуют элементы вектора y для минимального собственного числа с значимо отличающимися от 0 значениями и порождают мультиколлинеарность.

Вопрос 11

$$p_{ii} = f^T(x_i)(X^T X)^{-1} f(x_i)$$

. На практике руководствуются следующими правилами. Значения $p_{ii} \leq 0,2$ позволяют считать надежными данные наблюдения, значения $0,2 \leq p_{ii} \leq 0,5$ относят данное наблюдение к классу рискованных, а при значениях $p_{ii} > 0,5$ i -е наблюдение можно принять за точку разбалансировки.

Вопрос 12

В 4 раза.

Вопрос 13

Монотонность коэффициента детерминации от числа включенных в модель регрессоров легко доказывается, если обратиться к его выражению:

$$R^2 = 1 - \frac{RSS}{\sum (y_i - \bar{y})^2}$$

Вопрос 14

При наличии повторной выборки можно предложить критерий селекции, в основу которого положена идея идеального внешнего критерия. При вычислении идеального внешнего критерия вместо зашумленных значений отклика используются точные (незашумленные)

значения. Однако в реальной ситуации эти значения нам не известны. В случае повторной выборки вместо неизвестных значений вектора можно использовать и их оценку в виде выборочного среднего $\bar{y}$:

$$\bar{y}^T = (\bar{y}_A^T, \bar{y}_B^T), \bar{y}_{A_i} = (y_{A_i} + y_{B_i})/2, i = 1, \dots, n_A, \bar{y}_B = \bar{y}_A.$$

С использованием вектора $\bar{y}$ можно предложить следующий субидеальный критерий стабильности:

$$\tilde{S}^2 = \|\bar{y}_A - X_A \hat{\theta}_A\|^2 + \|\bar{y}_B - X_B \hat{\theta}_B\|^2.$$

Вопрос 15

С целью управления масштабом оценок введем в рассмотрение функцию стоимости

$$C = \sum_{i=1}^n (y_i - f(x_i)^T \theta)^2 + \sum_{j=1}^m \lambda_j \theta_j^2,$$

где второе слагаемое рассматривается как штраф при условии, что $\lambda_j \geq 0$. Решим данную оптимизационную задачу $C \rightarrow \min$.

Представим целевую функцию в матричном виде

$$C = (y - X\theta)^T (y - X\theta) + \theta^T \Lambda \theta,$$

где $\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_m \end{pmatrix}.$

Найдем решение из условия равенства нулю первых производных:

$$\frac{\partial C}{\partial \theta_j} = 2X^T y + 2X^T X \hat{\theta} + 2\Lambda \hat{\theta} = 0,$$

$$\hat{\theta} = (X^T X + \Lambda)^{-1} X^T y.$$

Ридж-оценки впервые были введены А. Гоэрлом при исследовании регрессии 2-го порядка (ridge – гребень, хребет). Название "ридж-оценки" навеяно следующими наблюдениями. При сильной мультиколлинеарности линии равного уровня доверительного эллипсоида для разного уровня значимости (доверительной вероятности) становятся сильно вытянутыми и визуально похожими на хребет на географической карте. Данные оценки использовались именно в условиях мультиколлинеарности и поэтому получили название ридж-оценок.

Понятно, что ридж-ценки существенно зависят от выбранной матрицы Λ , в силу чего имеют обозначение

$$\hat{\theta}(\Lambda) = (X^T X + \Lambda)^{-1} X^T y.$$

Часто матрицу Λ задают диагональной в виде $\Lambda_{ii} = \lambda \cdot (X^T X)_{ii}$, $\lambda \geq 0$, т.е. пропорциональной диагональным элементам матрицы $X^T X$. Иногда применяют ещё более простой вариант $\Lambda = \lambda \cdot I_m$. Можно показать, что к данному случаю сводится предыдущий.

Обозначим

$$D = \begin{pmatrix} (X^T X)_{11} & & 0 \\ & \ddots & \\ 0 & & (X^T X)_{mm} \end{pmatrix},$$

$$Z = XD^{-1/2}, \quad \beta = D^{1/2} \theta.$$

Тогда исходная модель

$$y = X\theta + \varepsilon$$

сводится в новых обозначениях к модели

$$y = Z\beta + \varepsilon = XD^{-1/2}D^{1/2}\theta + \varepsilon = X\theta + \varepsilon.$$

Ридж-оценка $\hat{\beta}(\lambda)$ с матрицей $\lambda \cdot I_m$ будет равна

$$\begin{aligned} \hat{\beta}(\lambda) &= (Z^T Z + \lambda I)^{-1} Z^T y = (D^{-1/2} X^T X D^{-1/2} + \lambda I)^{-1} D^{-1/2} X^T y = \\ &= (X^T X + \lambda D)^{-1} D D^{-1/2} X^T y = D^{1/2} (X^T X + \lambda I)^{-1} X^T y, \end{aligned}$$

где $\Lambda_{ii} = \lambda (X^T X)_{ii}$. Поэтому $\hat{\theta}(\Lambda) = D^{-1/2} \hat{\beta}(\lambda)$.

Итак, можно рассматривать более простой случай

$$\hat{\beta}(\lambda) = (Z^T Z + \lambda I)^{-1} Z^T y.$$

Свойства этой оценки:

- 1) $\hat{\beta}(0)$ - МНК-оценка $\hat{\beta}$;
- 2) $\hat{\beta}(\lambda)$ - является линейным преобразованием МНК-оценки и является смещённой;
Действительно,

$$Z^T y = (I + \lambda (Z^T Z)^{-1})^{-1} (Z^T Z)^{-1} Z^T y = B \hat{\beta},$$

где $B = (I + \lambda (Z^T Z)^{-1})^{-1}$ - матрица линейного преобразования и $B \neq I$, что приводит к смещению оценок $\hat{\beta}(\lambda)$ в отличие от $\hat{\beta}$, которые несмещены.

- 3) в классе оценок с фиксированной длиной ридж-оценка минимизирует сумму квадратов отклонений;
- 4) верно и обратное утверждение: ридж-оценка имеет минимальную длину в классе оценок с заданным значением суммы квадратов отклонений (RSS).

Доказано, что всегда существует такое значение параметра λ , что ССКО для оценки $\hat{\beta}(\lambda)$ будет меньше чем ССКО для оценки $\hat{\beta}(0)$. Показано также, что

$$\hat{\beta}^T(\lambda)\hat{\beta}(\lambda) = (\mu_{\max} / (\mu_{\max} + \lambda)) \hat{\beta}^T \hat{\beta},$$

где $\mu_{\max}$ - максимальное собственное число информационной матрицы $Z^T Z$. Это означает, что длина ридж-оценки будет уменьшаться с увеличением параметра регуляризации, но это и так понятно из рассмотрения целевой функции, где используется штраф за длину оценки.

Задача

1. Модель $\eta_1 = \theta_1 x + \theta_2 x^2$
2. Гипотеза принимается

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.

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

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: 02.03.03 Математическое обеспечение и администрирование информационных систем, профиль: Математическое и программное обеспечение информационных технологий

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 решать стандартные задачи профессиональной деятельности на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий и с учетом основных требований информационной безопасности	у5. уметь использовать специализированные программные средства при решении профессиональных задач	Интервальное оценивание неизвестных параметров регрессионных, моделей проверка гипотез 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	Экзамен, вопросы 1-109, задача
ОПК.2 способность применять в профессиональной деятельности знания математических основ информатики	з8. Знать теоретические основы методов построения зависимостей по экспериментальным данным	Критерии качества модели. Алгоритмы выбора наилучшей модели регрессии: алгоритмы включения и исключения, эффективный алгоритм построения всех регрессий, многорядный комбинаторный алгоритм Задачи статистического исследования зависимостей, задачи снижения размерности, задачи классификации, задачи кластеризации. Интервальное оценивание и проверка гипотез. Оценивание неизвестной дисперсии ошибки наблюдения. Проверка адекватности регрессионной модели. Основные требования, предъявляемые к оценкам параметров. Наилучшие линейные оценки (НЛО). Вычисление НЛО, их свойства. Связь НЛО с оценками по методу наименьших квадратов (МНК). Оценивание параметров в модели регрессии при наличии линейных ограничений. Проверка общих линейных гипотез Оценивание		Экзамен. Вопросы 1-12, 14-34, 51-109

		параметров в условиях гетероскедастичности, автокорреляции ошибок наблюдения, в условиях мультиколлинеарности		
ОПК.2	уб. Уметь использовать базовые знания алгебры, вычислительной математики, теории вероятностей при разработке вычислительных алгоритмов построения многофакторных моделей зависимостей по экспериментальным данным	Интервальное оценивание и проверка гипотез. Оценивание неизвестной дисперсии ошибки наблюдения. Проверка адекватности регрессионной модели. Основные требования, предъявляемые к оценкам параметров. Наилучшие линейные оценки (НЛО). Вычисление НЛО, их свойства. Связь НЛО с оценками по методу наименьших квадратов (МНК). Оценивание параметров в модели регрессии при наличии линейных ограничений. Проверка общих линейных гипотез		Экзамен. Вопросы 13-21, 51-109, задача
ПК.3/ПК готовность к разработке моделирующих алгоритмов и их реализации их на базе современных языков и программных систем	у4. Уметь реализовывать вычислительные алгоритмы построения зависимостей по экспериментальным данным	Интервальное оценивание неизвестных параметров регрессионных, моделей проверка гипотез 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	Экзамен, вопросы 1-109, задача
ПК.8.В/ПК готовность к использованию основных моделей информационных технологий для решения прикладных задач	у3. Владеть методами построения и анализа линейных моделей по экспериментальным данным	Интервальное оценивание неизвестных параметров регрессионных, моделей проверка гипотез 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	Экзамен, вопросы 1-109, задача

2. Методика оценки этапов формирования компетенций в рамках дисциплины.

Промежуточная аттестация по дисциплине проводится в 7 семестре - в форме экзамена, который направлен на оценку сформированности компетенций ОПК.1, ОПК.2, ПК.3/ПК, ПК.8.В/ПК. Время подготовки на экзамене 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).

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.