Министерство образования и науки РФ
Федеральное агентство по образованию
ГОУ ВПО «Амурский государственный университет»

УТВЕРЖДАЮ
Зав. кафедрой ОМиИ
_____Г.В.Литовка
«__» _____2008 г.

УЧЕБНО-МЕТОДИЧЕСКИЙ КОМПЛЕКС ПО ДИСЦИПЛИНЕ «Математические методы моделирования социальных процессов»

для специальности 040201 – «Социология»

Составитель: Н.Н. Двоерядкина, к.п.н.

Благовещенск, 2008

Печатается по решению

редакционно-издательского совета

факультета математики и информатики

Амурского государственного университета

Двоерядкина Н.Н.

Учебно-методический комплекс дисциплины «Математические методы моделирования социальных процессов» для специальности 040201 –

Благовещенск: АмГУ, 2008. – 50 с.

Содержание

1. Рабочая программа дисциплины.

1.1. Цели и задачи дисциплины, ее место в учебном процессе.

Программа предназначена для подготовки дипломированных специалистов по специальности «Социология». Это накладывает на нее определенные особенности, заключающиеся в том, что выпускник должен получить высшее образование, способствующее дальнейшему развитию личности.

Дисциплина «Математическое моделирование социальных процессов» является составной частью математического образования социолога, логичным продолжением курсов «Математики» и «Математических методов в социологии». Она призвана обеспечить специалиста-социолога мощным средством решения прикладных профессиональных задач.

Математическое моделирование социальных процессов является междисциплинарным инструментарием, который выполняет две основные функции: первую — обучающую специалиста-профессионала умению правильно задавать цель тому или иному процессу, определить условия и ограничения в достижении цели; вторую – аналитическую, т.е. «проигрывание» на моделях возможных ситуаций и получение оптимальных решений.

Именно поэтому при обучении социологов математическому моделированию социальных процессов основное внимание уделяется не строгим математическим доказательствам того или иного утверждения, а построению и анализу моделей конкретных жизненных ситуаций.

Благодаря междисциплинарным связям, осуществляемым в процессе преподавания дисциплины и моделированию социальной действительности студенты начинают понимать какое место математика и математические методы занимают в их профессиональной деятельности.Основной целью данной дисциплины, в связи с постоянно усложняющимися процессами, требующими создания и совершенствования особых методов изучения и

анализа, является обучение студентов практическим навыкам составления математических моделей.

Для достижения поставленной цели в рамках дисциплины решаются следующие задачи:

1) Анализ известных моделей с позиций их устойчивости к незначительным изменениям окружающей действительности.

2) Сбор и обработка «сырой» информации, необходимой для количественной и качественной оценки социальных процессов.

3) Построение моделей путем анализа первичной информации и нахождение параметров модели.

Дисциплина изучается после освоения студентами:

- высшей математики;
- теории вероятностей и математической статистики;
- экономико-математических методов;
- эконометрики.

1.2.Содержание дисциплины.

Дисциплина «Математическое моделирование социальных процессов» является дисциплиной по выбору, изучается в 7 семестре. Учебным планом специальности «Социолоия» на ее изучение отводится 100 часов, в том числе 28 ч – лекций, 14 ч – практические занятия, 58 ч – самостоятельная работа.

Лекционные занятия, наименование тем, содержание, объем в часах.

1. Дискриминантный анализ: дискриминантные функций, оценка их надежности, построение классифицирующих функций, классификационная матрица, расстояния Махаланобиса, апостериорные и априорные вероятности (6 ч).

2. Многомерное шкалирование: построение геометрического образа социального пространства, метрические и неметрические методы шкалирования, показатель «стресса» (4 ч).

3. Регрессионный анализ: модель множественной регрессии, фиктивные переменные и их использование в моделях, логистические и лог-линейные модели, их анализ. Устойчивость моделей к изменениям внутри системы и внешней среды (6 ч).

4. Канонические корреляции: канонические величины генеральной совокупности и их интерпретация, оценка канонических корреляций (4 ч).

5. Анализ временных рядов: временные ряды при изучении динамики социальных явлений, тренд, сезонность, построение моделей с аддитивной и мультипликативной сезонностью (4 ч).

6. Робастные методы оценивания: грубые ошибки и методы их выявления в статистической совокупности данных, критерии обнаружения и исключения грубых ошибок, методы вычисления устойчивых статистических оценок (4 ч).

Практические занятия, их содержание и объем в часах

1. Дискриминантный анализ : нахождение дискриминантных функций, использование χ^2-критерия для определения оптимального количества дискриминантных функций, классификация наблюдений при помощи дискриминантных функций, классифицирующих функций, расстояния Махаланобиса, апостериорных вероятностей (4 ч).

2. Многомерное шкалирование: построение геометрического образа социального пространства, метрические и неметрические методы шкалирования, вычисление показателя «стресса» (2 ч).

3. Регрессионный анализ: анализ и построение моделей множественной регрессии, логистических и лог-линейных моделей, анализ фиктивных переменных в моделях (2 ч).

4. Анализ временных рядов: построение моделей временного ряда с аддитивной и мультипликативной сезонностью. (4 ч)

5. Робастные методы оценивания: методы выявления грубых ошибок в статистической совокупности данных, критерии обнаружения и исключения грубых ошибок, методы вычисления устойчивых оценок (2 ч).

1.3. Распределение времени по курсу.

№ недели	Вопросы, изучаемые на лекции	Вопросы, изучаемые на практическом занятии	Самостоятельная работа	Формы контроля
1	Дискриминантные функций, оценка их надежности		Выполнение комплексного задания	Защита задания
				к/р
2	Построение классифицирующих функций, расстояния Махаланобиса.	Нахождение дискриминантных функций, χ^2-критерия		
3	Апостериорные и априорные вероятности			
4	Метрические методы шкалирования	Классификация наблюдений при помощи дискриминантных классифицирующих функций, , апостериорных вероятностей.		
5	Неметрические методы шкалирования,			
6	Множественной регрессии с фиктивными переменными	Построение геометрического образа социального пространства		
7	Логистические модели, их анализ			
8	Лог-линейные модели, их анализ	Построение различных регрессионных моделей и их анализ.		
9	Временные ряды при изучении динамики социальных явлений, тренд.			
10	Сезонность при исследовании временных рядов	Построение моделей временного ряда с аддитивной сезонностью.		
11	Канонические величины генеральной совокупности и их интерпретация			
12	Оценка канонических корреляций	Построение моделей временного ряда с мультипликативной сезонностью.		
13	Грубые ошибки, критерии обнаружения и исключения грубых ошибок			
14	Методы вычисления устойчивых статистических оценок	Исключения грубых ошибок, вычисление устойчивых статистических оценок.		

1.4. Примерные вопросы к зачету.

1. Понятие о многомерных статистических методах исследования.
2. Границы применимости многомерных статистических методов.
3. Классификация многомерных статистических методов.
4. Примеры задач, решаемых с помощью многомерных статистических методов исследования.
5. Построение математических моделей различных социально-экономических задач, основные этапы.
6. Моделирование значений наблюдаемых переменных на основе фиктивных факторов.
7. Постановка задачи многомерного шкалирования.
8. Построение геометрического образа социального пространства.
9. Метрические и неметрические методы шкалирования, показатель «стресса».
10. Постановка задачи дискриминантного анализа.
11. Понятие о дискриминантных функциях.
12. Определение оптимального количества дискриминантных функций.
13. Оценка параметров дискриминантных функций и их качества.
14. Классификация объектов и наблюдений при помощи дискриминантных функций.
15. Построение классифицирующих функций, классификация объектов и наблюдений при помощи классифицирующих функций.
16. Вычисление расстояний Махаланобиса, апостериорных расстояний и классификация объектов с их помощью.
17. Понятие классификационной матрицы, ее анализ.
18. Временные ряды при изучении динамики социальных явлений.
19. Понятие тренда, сезонности, аддитивная и мультипликативная сезонность.
20. Примеры временных рядов с аддитивной и мультипликативной сезонностью.

21. Построение регрессионных моделей
22. Модели логистической регрессии и лог-линейной регрессии.
23. Системы регрессионных уравнений. Примеры моделей с использованием систем уравнений.
24. Идентифицируемость и сверхидентифицируемость систем.
25. Эндогенные и экзогенные переменные.
26. Анализ устойчивости моделей к изменениям внутри системы и внешней среды.
27. Понятие грубых ошибок.
28. Причины появления ошибок в статистической совокупности.
29. Существующие подходы при обработке грубых ошибок.
30. Основные методы устойчивого оценивания патаметров выборочной совокупности.
31. Отличие канонических корреляций от корреляционно-регрессионного анализа.
32. Понятие канонических переменных.
33. Приемы канонического анализа для сокращения количества исследуемых факторов.

1.5. Рекомендуемая литература.

1. Толстова Ю.Н. Основы многомерного шкалирования– М.: Книжный дом Университет,2006. –158с.
2. Плотинский Ю.М.Теоретические и эмпирические модели социальных процессов: учеб. пособие для вузов. – М.:Логос, 1998. -280 с.
3. СамарскийА.А.Математическое моделирование: идеи, методы, проблемы: Моногр./А.А.Самарский,А.П.Михайлов.– М.: 2005. – 320 с.

4. Советов, Борис Яковлевич. Моделирование систем [Текст] : учеб.: рек. Мин. обр. РФ / Б. Я. Советов, С. А. Яковлев. - 4-е изд., стер. - М. : Высш. шк., 2005. - 344 с
5. Бородкин, Фридрих Маркович. Социальные индикаторы [Текст] : учеб.: рек. УМО / Ф. М. Бородкин, С. А. Айвазян. - М. : ЮНИТИ-ДАНА, 2006. - 608 с.
6. Лбов,Г.С. Логические решающие функции и вопросы статистической устойчивости решений [Текст]: научное издание /Г.С. Лбов, Н.Г. Старцева. - Новосибирск : Изд-во Ин-та мат-ки, 1999. - 212 с.
7. Мжельский,Б.И. Математические модели задач оптимизации [Текст] : сб. задач / Мжельский Б.И., Мжельская В.А. - М. : Изд-во МЭИ, 1998. - 64с.

2. МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ И УКАЗАНИЯ.

2.1. Методические рекомендации по проведению лекций

Для математики важна не природа рассматриваемых объектов, а лишь существующие между ними соотношения, в то время как социолог работает в основном с реальными данными.

Курс «Математическое моделирование социальных процессов» призван размыть грань между формализмом математики и реальными происходящими в обществе процессами, интересующими социолога.

Значение лекционных занятий по данному курсу обусловлено следующими причинами:

- отсутствием единого учебника, в котором изложены всевозможные математические методы, используемые в социологии;
- необходимостью адаптировать лектором некоторые математические методы для нужд социолога;
- невозможностью студента самостоятельно представить смысл полученных им знаний по математике.

Каждая лекция сопровождается высоким научным стилем изложения и достаточным количеством примеров профессионального характера, которые разрешают противоречие между желанием поскорее приобщиться к профессии и необходимостью терпеливого изучения фундаментальных дисциплин.

Лекция 1 носит обзорно-повторительный характер. На ней происходит знакомство студентов с целью, назначением и местом курса в системе учебных дисциплин. Большая часть лекции отводится на изучение основ дискриминантного анализа. Вводится понятие дискриминантных функций и осуществляется классификация объектов с помощью этих функций

На второй и третьей лекциях продолжается знакомство с методами дискриминантного анализа. Вводится понятие классифицирующих функций, апостериорной и априорной вероятностей и методов их расчета.

Метрическим и неметрическим методам многомерного шкалирования отводятся 4 и 5 лекции курса. Рекомендуется указать сходства и различия данных методов, охарактеризовать показатель стресса и рассмотреть построение геометрического образа социального пространства.

При изложении лекций, посвященных регрессионному анализу уделяется большое внимание построению моделей множественной регрессии, так как большинство характеристик, исследуемых в социологии зависят от нескольких переменных величин. Кроме того, необходимо рассмотреть модели, в состав которых входят фиктивные переменные (пол, образование и др.).

Большая часть данных в современном мире имеют вид временных рядов. Начиная с 1960-х гг. временные ряды стали регулярно использовать для прогнозов, а на сегодняшний день потребность в прогнозах только усиливается. Поэтому при изучении математического моделирования социальных процессов несколько занятий посвящается анализу временных рядов.

Краткий конспект лекций по каждой теме приводится в п.3.1.

2.2. Методические рекомендации к практическим занятиям.

Лекционный курс дисциплины «Математика в экономике» сопровождается практическими занятиями. Теоретические знания, представления, образы должны быть прожиты. Афоризм одного из известных физиков М. Лауэ: знание есть то, что остается, когда все выученное уже забыто, характеризует важную роль практики.

Практические занятия должны проводиться в логичном единстве с теоретическим курсом, подкрепляя и уточняя понятийный аппарат, путем решения задач профессиональной направленности.

Каждый практическое занятие начинается с теоретического опроса необходимого материала и проверки домашнего задания. Далее на конкретных примерах рассматриваются пути и способы применения тех

математических методов, которые не требуют использования электронных вычислительных машин. При этом необходимо активизировать самостоятельную работу студентов. Задания и методические указания к ним выдаются студентам, каждый из которых выбирает оптимальный для себя темп работы. Преподавателю отводится роль консультанта и помощника. Задания, вызвавшие трудности у большинства студентов, разбираются на доске.

При работе студенты должны опираться на систему базовых математических знаний, приобретенных при изучении высшей математики, теории вероятностей и математической статистики, и понимать качественный смысл тех количественных преобразований, которые они осуществляют с помощью математических методов и моделей.

В конце занятия выдается домашнее задание, состоящее из теоретических вопросов, уяснение которых необходимо для следующего занятия и практических заданий по пройденному материалу.

2.3. Методические указания по выполнению домашних заданий.

При выполнении домашнего задания решать задачи удобнее поэтапно, в той последовательности, в какой эти задания сформулированы. В этом случае при возникновении трудностей будет легче обратиться к анализу тех тем, которые изложены в лекции и задач, разобранных на практическом занятии.

Следует иметь в виду, что решение задач направлено на выработку навыков моделирования. Поэтому при выполнении заданий требуется абстрагироваться от содержательного анализа предлагаемых задач и формально применить необходимый математический аппарат

При выполнении заданий ответы должны быть аргументированными, то есть недостаточно просто привести ответ, необходимо указать путь, каким Вы пришли к данному ответу, и те основания, которыми Вы руководствовались. При этом следует обратить внимание на то, что ряд заданий предусматривает несколько последовательных шагов или операций

для ответа на вопрос. При получении ответа в задаче необходимо правильно интерпретировать его, согласно условию, даже если на Ваш взгляд, данный результат не соответствует действительности.

В случае затруднения с определением алгоритма, необходимого для решения конкретных задач, а также типового оформления ответа на задание, рекомендуется обратиться к образцам выполнения типичных задач, которые представлены на практическом занятии.

После выполнения практической части задания следует найти ответы на теоретические вопросы, заданные преподавателем и таким образом подготовится к осознанному восприятию следующего материала.

Активная, регулярная самостоятельная работа над домашним заданием – путь к успешному усвоению дисциплины.

2.4. Методические указания по выполнению контрольных работ.

По курсу «Математическое моделирование социальных процессов» предусмотрена одна итоговая контрольная работа. Целью контрольной работы является выявление уровня знаний студентов и умений определять виды задач, к которым применимы многомерные методы.

Написание контрольной работы формирует у студентов способность абстрагироваться от фабулы задачи, строить формализованную математическую модель предложенных явлений, выделять общие закономерности и особенности многомерных методов исследования.

При подготовке к контрольной работе студенту необходимо изучить и систематизировать теоретический материал по теме. Разобрать конкретные примеры. Решить достаточное количество задач и упражнений, во время аудиторной и самостоятельной домашней работы.

2.5. Методические указания по организации контроля знаний студентов.

Основной целью учебного процесса в вузе является подготовка высококвалифицированных специалистов, способных творчески решать

профессиональные задачи. Контроль и оценка знаний умений и навыков является одним из важных аспектов обучения, который существенно влияет на его качество.

Контролю знаний присущи определенные дидактические правила: объективность, действенность, систематичность, индивидуальность, единство требований.

Отчет по материалу курса только на зачете не может обеспечить полноту его усвоения студентами. Поэтому в течение семестра предусмотрены и другие виды контроля. При преподавании дисциплины «Математическое моделирование социальных процессов» используются три основных вида контроля знаний студентов – текущий, тематический и итоговый.

При текущем контроле оценивается уровень участия студентов в аудиторной работе, степень усвоения ими учебного материала и выявляются недочеты по подготовке студентов в целях дальнейшего совершенствования методики преподавания данной дисциплины, активизации работы студентов в ходе занятия и оказания им индивидуальной помощи.

Текущий контроль проводится непосредственно на лекциях, и практических занятиях. В процессе чтения лекций преподаватель работает с аудиторией и по ее реакции оценивает степень усвоения материала. В ходе или в конце лекции студентам задается несколько вопросов по изложенной теме, что способствует закреплению полученных знаний. На практических занятиях текущий контроль проводится индивидуально. Полученные знания и степень усвоения материала проверяются в устной или письменной форме.

Тематический контроль проводится после прохождения крупных тем или разделов.

Итоговым контролем является зачет. Успешная сдача зачета обусловлена знанием теории, умением решать практические задачи и составлять математические модели.

3. КОМПЛЕКТЫ ЗАДАНИЙ К ЗАНЯТИЯМ

3.1. Краткий конспект лекций.

Введение. Многомерный анализ для исследования социальных процессов.

Социальные процессы и явления зависят от большого числа параметров, их характеризующих, что обусловливает трудности, связанные с выявлением структуры взаимосвязей этих параметров. В подобных ситуациях, т. е. когда решения принимаются на основании анализа стохастической, неполной информации, использование методов многомерного статистического анализа является не только оправданным, но и существенно необходимым.

Многомерные статистические методы среди множества возможных вероятностно-статистических моделей позволяют обоснованно выбрать ту, которая наилучшим образом соответствует исходным статистическим данным, характеризующий реальное поведение исследуемой совокупности объектов, оценить надежность и точность выводов, сделанных на основании ограниченного статистического материала.

К области приложения многомерных статистических методов могут быть отнесены задачи, связанные с исследованием поведения индивидуума, семьи или другой социально-экономической или производственной единицы, как представителя большой совокупности объектов.

Выделяют три центральные задачи, решаемые с помощью многомерных методов.

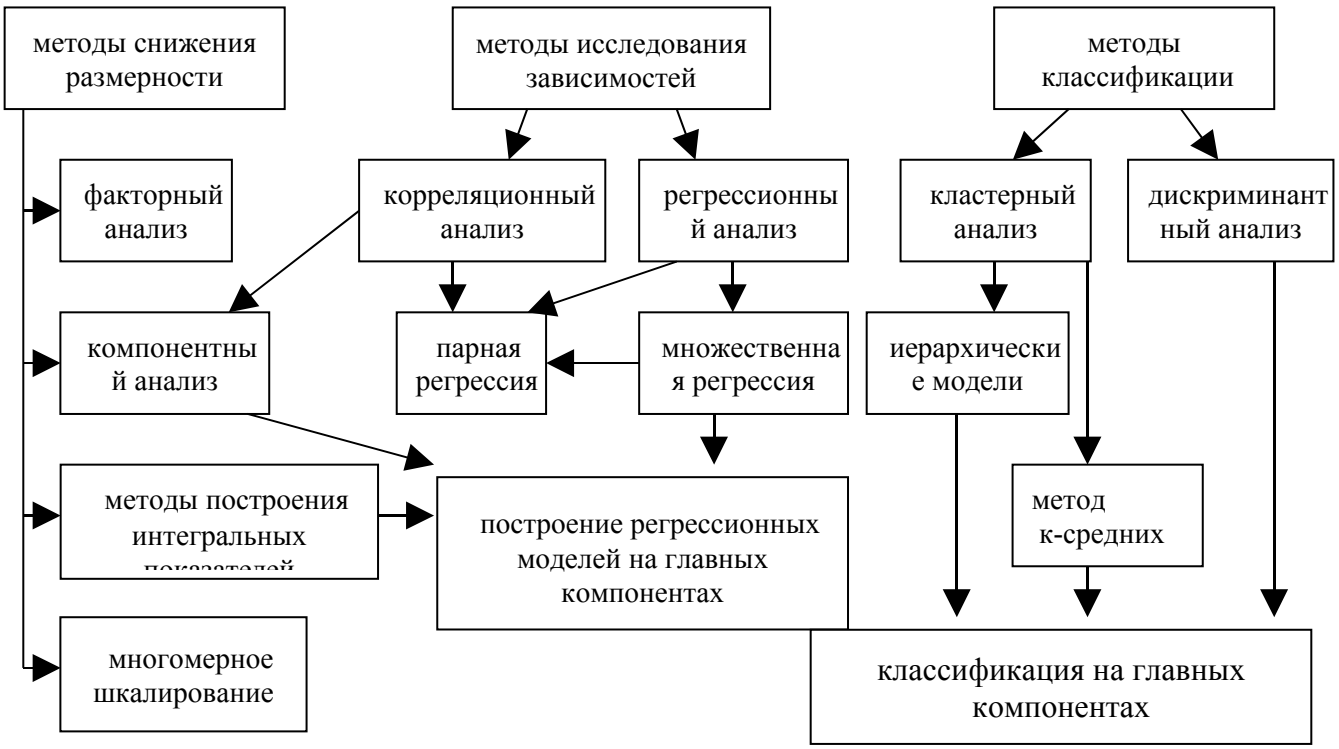
1. Статистическое исследование структуры и характера взаимосвязей, существующих между анализируемыми количественными переменными. При этом под переменными понимаются как регистрируемые на объектах признаки, так и время *t*.

2. Разработка статистических методов классификации объектов и признаков.

3. Снижение размерности исследуемого признакового пространства с целью лаконичного объяснения природы анализируемых многомерных данных. Возможность лаконичного описания анализируемых многомерных данных основана на априорном допущении, в соответствии с которым существует небольшое число признаков с помощью которых могут быть достаточно точно описаны как сами наблюдаемые переменные анализируемых объектов, так и определяемые этими переменными свойства (характеристики) анализируемой совокупности. При этом упомянутые признаки могут находиться среди исходных признаков, а могут быть латентными, т.е. непосредственно статистически не наблюдаемыми, но восстанавливаемыми по исходным данным.

Эти задачи не исчерпывают всех возможностей многомерных статистических методов, но в настоящий момент являются наиболее распространенными.В соответствии с задачами в структуре многомерных статистических методов выделяют методы снижения размерности, методы исследования зависимостей, методы классификации.

Классификацию многомерных методов представим на схеме:



Работая с многомерными статистическими методами важно, чтобы переменные изменялись в сравнимых шкалах. Из неоднородности единиц измерения вытекает невозможность обоснованного выражения значений различных показателей в одном масштабе. Чтобы устранить неоднородность измерения исходных данных, все их значения предварительно нормируются, т.е. выражаются через отношение этих значений к некоторой величине, отражающей определенные свойства данного показателя. Нормирование исходных данных иногда проводится посредством деления исходных величин на среднеквадратичное отклонение соответствующих показателей. Другой способ сводится к вычислению, так называемого, стандартизованного вклада. Его еще называют *Z*-вкладом.

Z - вклад показывает, сколько стандартных отклонений отделяет данное наблюдение от среднего значения:

$Z_i = \frac{x_i - \bar{x}}{\sigma_i}$, где x_i - значение данного наблюдения, $\bar{x}$ - среднее, σ_i -стандартное отклонение.

Среднее для *Z*-вкладов является нулевым и стандартное отклонение равно 1.

Стандартизация позволяет сравнивать наблюдения из различных распределений.

Заметим, что методы нормирования означают признание всех признаков равноценными с точки зрения выяснения сходства рассматриваемых объектов. Признание равноценности различных показателей кажется оправданным отнюдь не всегда. Было бы желательным наряду с нормированием придать каждому из показателей вес, отражающий его значимость в ходе установления сходств и различий объектов.

В этой ситуации приходится прибегать к способу определения весов отдельных показателей – опросу экспертов.

Экспертные оценки дают известное основание для определения важности индикаторов, входящих в ту или иную группу показателей.

Довольно часто при решении подобных задач используют не один, а два расчета: первый, в котором все признаки считаются равнозначными, второй, где им придаются различные веса в соответствии со средними значениями экспертных оценок.

Тема 1. Дискриминантный анализ

Дискриминантный анализ является разделом многомерного статистического анализа, который позволяет изучать различия между двумя и более группами объектов по нескольким переменным одновременно.

При интерпретации нужно ответить на вопрос: возможно ли, используя данный набор переменных, отличить одну группу от другой, насколько хорошо эти переменные помогают провести дискриминацию и какие из них наиболее информативны.

Методы классификации связаны с получением одной или нескольких функций, обеспечивающих возможность отнесения данного объекта к одной из групп. Эти функции называются классифицирующими и зависят от значений переменных таким образом, что появляется возможность отнести каждый объект к одной из групп.

Задачи дискриминантного анализа можно разделить на три типа. Задачи первого типа позволяют на основе некоторой информации найти функцию, позволяющую поставить в соответствие новым объектам характерные для нихвыводы. Построение такой функции и составляет задачу дискриминации. Второй тип задачи относится к ситуации, когда признаки принадлежности объекта к той или иной группе потеряны, и их нужно восстановить. Задачи третьего типа связаны с предсказанием будущих событий на основании имеющихся данных.

Дискриминантный анализ общий термин, относящийся к нескольким тесно связанным статистическим процедурам. Эти процедуры можно разделить на методы интерпретации межгрупповых различий – *дискриминации* и методы *классификации* наблюдений по группам.

Основной целью дискриминации является нахождение такой линейной комбинации переменных (в дальнейшем эти переменные будем называть дискриминантными переменными), которая бы оптимально разделила рассматриваемые группы.

Введем следующие обозначения:

g – число классов;

p – число дискриминантных переменных;

n_k – число наблюдений в *k*–й группе;

n – общее число наблюдений по всем группам;

x_{ikm} – величина дискриминантной переменной *i* для *m*–го наблюдения в *k*–й группе;

$\bar{x}_{ik}$ – средняя величина переменной *i* в *k*–й группе;

$\bar{x}_i$ – среднее значение переменной i по всем группам;

Линейная функция $d_{km} = \beta_0 + \beta_1 x_{1km} + \ldots + \beta_p x_{pkm}$, $m = 1,\ldots n$; $k = 1,\ldots g$ называется *канонической дискриминантной функцией* с неизвестными коэффициентами β_i, d_{km} –значение дискриминантной функции для *m*–го объекта в группе *k*

Коэффициенты β_i первой канонической дискриминантной функции выбираются таким образом, чтобы центроиды различных групп как можно больше отличались друг от друга. Коэффициенты второй группы выбираются также, но при этом налагается дополнительное условие, чтобы значения второй функции были некоррелированы со значениями первой. Аналогично определяются и другие функции. Отсюда следует, что любая каноническая дискриминантная функция имеет нулевую внутригрупповую корреляцию с остальными. Если число групп равно *g*, то число канонических дискриминантных функций будет на единицу меньше числа групп. С геометрической точки зрения дискриминантные функции определяют гиперповерхности в *p*-мерном пространстве. В частном случае при *p*=2 она является прямой, а при *p*=3 плоскостью

На практике полезно иметь одну, две или же три дискриминантных функций. Тогда графическое изображениее объектов будет представлено в одно–, двух– и трехмерных пространствах. Такое представление особенно полезно в случае, когда число дискриминантных переменных p велико по сравнению с числом групп g.

Для получения коэффициентов β_i канонической дискриминантной функции нужен статистический критерий различения групп. Классификация переменных будет осуществляться тем лучше, чем меньше рассеяние точек относительно центроида внутри группы и чем больше расстояние между центроидами групп. Один из методов поиска наилучшей дискриминации данных заключается в нахождении такой канонической дискриминантной функции d, которая бы максимизировала отношение межгрупповой вариации к внутригрупповой .

$$\lambda = \frac{B}{W}$$

где B – межгрупповая, а W – внутригрупповая матрицы рассеяния наблюдаемых переменных от средних. Иногда вместо W используют матрицу рассеяния T объединенных данных.

Рассмотрим задачу максимизации отношения для λ когда имеются g групп. Оценим информацию, характеризующую степень различия между объектами по всему пространству точек, определяемому переменными групп. Для этого вычислим матрицу рассеяния T, которая равна сумме квадратов отклонений и попарных произведений наблюдений от общих средних по каждой переменной. Элементы матрицы T определяются выражением

$$t_{ij} = \sum_{k=1}^{g} \sum_{m=1}^{n} (x_{ikm} - \bar{x}_i)(x_{jkm} - \bar{x}_j) \qquad \bar{x}_i = (1/n) \sum_{k=1}^{g} n_i \bar{x}_{ik}, i = 1, \ldots, p$$

$$\bar{x}_{ik} = (1/n_i) \sum_{m=1}^{n_k} x_{ikm}, i = 1, \ldots, p; k = 1, \ldots, g$$

Матрица T содержит полную информацию о распределении точек по пространству переменных. Диагональные элементы представляют собой

сумму квадратов отклонений от общего среднего и показывают, как ведут себя наблюдения по отдельно взятой переменной. Внедиагональные элементы равны сумме произведений отклонений по одной переменной на отклонения по другой.

Если разделить матрицу T на (n-1), то получим ковариационную матрицу. Для проверки условия линейной независимости переменных полезно рассмотреть вместо T нормированную корреляционную матрицу.

Для измерения степени разброса объектов внутри групп рассмотрим матрицу W, которая отличается от T только тем, что ее элементы определяются векторами средних для отдельных групп, а не вектором средних для общих данных. Элементы внутригруппового рассеяния определятся выражением

$$w_{ij} = \sum_{k=1}^{g} \sum_{m=1}^{n_k} (x_{ikm} - \bar{x}_{ik})(x_{jkm} - \bar{x}_{jk})$$

Если разделить каждый элемент матрицы W на (n–g), то получим оценку ковариационной матрицы внутригрупповых данных.

Когда центроиды различных групп совпадают, то элементы матриц T и W будут равны. Если же центроиды групп различные, то разница B=T-W будет определять межгрупповую сумму квадратов отклонений и попарных произведений. Если расположение групп в пространстве различается (т.е. их центроиды не совпадают), то степень разброса наблюдений внутри групп будет меньше межгруппового разброса.

Матрицы W и B содержат всю основную информацию о зависимости внутри групп и между группами. Для лучшего разделения наблюдений на группы нужно подобрать коэффициенты дискриминантной функции из условия максимизации отношения межгрупповой матрицы рассеяния к внутригрупповой матрице рассеяния при условии ортогональности дискриминантных плоскостей. Тогда нахождение коэффициентов

дискриминантных функций сводится к решению задачи о собственных значениях и векторах

Пусть $\lambda_1 \geq \ldots \lambda_p$ и $v_1 \ldots v_p$ соответственно собственные значения и векторы. Тогда $\lambda = \frac{\sum_k b_{jk} v_j v_k}{\sum_k w_{jk} v_j v_k}$

что влечет равенство $\sum_k (b_{jk} - \lambda w_{jk}) v_k = 0$, или в матричной записи $(B - \lambda W)v = 0$

Решение уравнения $|B - \lambda W| = 0$ позволяет нам определить компоненты собственных векторов, соответствующих дискриминантным функциям.

Каждое решение, которое имеет свое собственное значение и собственный вектор, соответствует одной дискриминантной функции. Координаты собственного вектора можно использовать в качестве коэффициентов дискриминантной функции. Однако при таком подходе начало координат не будет совпадать с главным центроидом. Для того, чтобы начало координат совпало с главным центроидом нужно нормировать координаты собственного вектора.

$$\beta_i = \mathbf{v}_i \sqrt{n - g}, \quad \beta_0 = -\sum_{i=1}^{p} \beta_i \bar{x}_i .$$

Нормированные коэффициенты полученые по нестандартизованным исходным данным, называются *нестандартизованными.* Нормированные коэффициенты приводят к таким дискриминантным значениям, единицей измерения которых является стандартное квадратичное отклонение. При таком подходе каждая ось в преобразованном пространстве сжимается или растягивается таким образом, что соответствующее дискриминантное значение для данного объекта представляет собой число стандартных отклонений точки от главного центроида.

Стандартизованные коэффициенты можно получить двумя способами:

1. стандартизировать исходные данные;

2. преобразовать нестандартизованные коэффициенты к стандартизованной форме по формуле:

$$\mathbf{c}_i = \boldsymbol{\beta}_i \sqrt{\frac{w_{ii}}{n-g}},$$

Стандартизованные коэффициенты полезно применять для уменьшения размерности исходного признакового пространства переменных. Если абсолютная величина коэффициента для данной переменной для всех дискриминантных функций мала, то эту переменную можно исключить, тем самым сократив число переменных.

Стандартизованные коэффициенты показывают вклад переменных в значение дискриминантной функции. Если две переменные сильно коррелированы, то их стандартизованные коэффициенты могут быть меньше по сравнению с теми случаями, когда используется только одна из этих переменных. Такое распределение величины стандартизованного коэффициента объясняется тем, что при их вычислении учитывается влияние всех переменных.

Общее число дискриминантных функций не превышает числа дискриминантных переменных и, по крайней мере, на 1 меньше числа групп. Степень разделения выборочных групп зависит от величины собственных чисел: чем больше собственное число, тем сильнее разделение. Наибольшей разделительной способностью обладает первая дискриминантная функция, соответствующая наибольшему собственному числу вторая обеспечивает максимальное различение после первой и т.д. Различительную способность *i*–ой функции оценивают по относительной величине в процентах собственного числа от суммы всех собственных чисел.

Коэффициент канонической корреляции. Другой характеристикой, позволяющей оценить полезность дискриминантной функции является коэффициент канонической корреляции r_i. Каноническая корреляция является мерой связи между двумя множествами переменных. Максимальная величина этого коэффициента равна 1. Будем считать, что группы

составляют одно множество, а другое множество образуют дискриминантные переменные. Коэффициент канонической корреляции для *i*–ой дискриминантной функции определяется формулой:

$r_i = \sqrt{\frac{\lambda_i}{1+\lambda_i}}$. Чем больше величина r_i, тем лучше разделительная способность дискриминантной функции.

После получения канонических дискриминантных функций при известной принадлежности объектов к тому или иному классу. решается задача предсказания класса, которому принадлежит некоторый случайно выбранный объект, т.е. задача классификации.

Классификация может проводиться с помощью апостериорной вероятности по Бейесовской схеме вычисления; с применением элементарных классифицирующих функций; с помощью расстояния Махалонобиса.

Классифицирующие функции имеют вид:

$$d_{ik} = b_{k0} + b_{k1}x_{i1} + \ldots + b_{kp}x_{ip} + \ln q_k,\ k = 1,\ldots,g$$

Объект относится к классу, у которого значение функции оказывается наибольшим. Коэффициенты классифицирующих функций удобнее вычислять по скалярным выражениям

$$b_{ki} = (n-g)\sum_{j=1}^{p}(w^{-1})_{ij}\bar{x}_{jk},\ k = 1,\ldots,g$$

где b_{ki} – коэффициент для переменной i в выражении, соответствующему классу k,

$(w^{-1})_{ij}$ – обратный элемент внутригрупповой матрицы сумм попарных произведений W. Постоянный член находится по формуле

$$b_{k0} = -0{,}5 \cdot \sum_{j=1}^{p} b_{kj}\bar{x}_{jk},\ k = 1,\ldots,g$$

Выбор функций расстояния между объектами для классификации является наиболее очевидным способом введения меры сходства для векторов объектов, которые интерпретируются как точки в евклидовом

пространстве. В качестве меры сходства можно использовать евклидово расстояние между объектами. Чем меньше расстояние между объектами, тем больше сходство. Однако в тех случаях, когда переменные коррелированы, измерены в разных единицах и имеют различные стандартные отклонения, трудно четко определить понятие "расстояния". В этом случае полезнее применить не евклидовое расстояние, а *выборочное расстояние Махаланобиса*

$$D^2(\mathbf{x}/G_k) = (n-g)\cdot\sum_{v=1}^{p}\sum_{j=1}^{p}(w^{-1})_{vj}(x_{iv}-\bar{x}_{vk})(x_{ij}-\bar{x}_{jk}),\ k=1,\ldots,g$$

При использовании функции расстояния, объект относят к той группе, для которой расстояние наименьшее.

Тема 2. Многомерное шкалирование.

Многомерное шкалирование - одно из направлений анализа данных; оно отличается от других методов многомерного статистического анализа, прежде всего видом исходных данных, которые в данном случае представляют собой матрицу близости между парами объектов («близость», или «сходство», объектов можно определять различными способами). Цель многомерного шкалирования - это описание матрицы близости в терминах расстояний между точками, представление данных о сходстве объектов в виде системы точек в пространстве малой размерности (например, на двумерной плоскости). Упрощая, можно сказать, что «на входе» методов многомерного шкалирования подается матрица близости, а «на выходе» получается координатное размещение точек.

Рассмотрим основные методические аспекты многомерного шкалирования.

Основное предположение многомерного шкалирования заключается в том, что существует некоторое метрическое пространство существенных базовых характеристик, которые неявно и послужили основой для полученных эмпирических данных о близости между парами объектов. Следовательно, объекты можно представить как точки в этом пространстве.

Предполагают также, что более близким (по исходной матрице) объектам соответствуют меньшие расстояния в пространстве базовых характеристик. Таким образом, многомерное шкалирование - это совокупность методов анализа эмпирических данных о близости объектов, с помощью которых определяется размерность пространства существенных для данной содержательной задачи характеристик измеряемых объектов и конструируется конфигурация точек (объектов) в этом пространстве. Это пространство («многомерная шкала») аналогично обычно используемым шкалам в том смысле, что значениям существенных характеристик измеряемых объектов соответствуют определенные позиции на осях пространства.

Данные в исходной матрице близости объектов могут быть получены различными способами. Методы многомерного шкалирования ориентируются на экспертные оценки близости объектов, когда респонденту предъявляют пары объектов, и он должен упорядочить их по степени внутреннего сходства, которое иногда оценивается в баллах. Если данные о близости пар объектов не получены непосредственно, а рассчитаны на основании других данных (различные коэффициенты связи), то следует иметь в виду, что многомерное шкалирование может оказаться далеко не лучшим способом анализа структуры исходных данных. Первичные данные, на основе которых рассчитывались близости, содержат больше информации, чем «вторичные» данные о близости.

Методы многомерного шкалирования делятся обычно на две категории: неметрическое многомерное шкалирование и метрическое многомерное шкалирование.

Методы метрического многомерного шкалирования используют, когда оценки близости получены на количественной шкале (не ниже интервальной). В таком виде в исследованиях социальных проблем оценки близости возникают крайне редко. Более естественной является оценка близости, измеренная на порядковой шкале (когда пары объектов можно

только упорядочить по степени схожести объектов). В этом случае используют методы неметрического многомерного шкалирования, которые дают «покоординатную развертку» матрицы близости в пространстве двух-трех существенных характеристик, так что упорядочения объектов по матрице близости расстояниям в этом пространстве совпадают.

Основные возможности методов многомерного шкалирования:

1. Построение метрического пространства невысокой размерности, в котором наилучшим образом сохраняется структура исходных данных о близости пар объектов. Проектирование объектов на оси полученного пространства определяет их положение на этих осях, т.е. производится процесс шкалирования.

2. Визуализация структуры исходных данных в виде кон фигурации точек (объектов) в двух-трехмерном базовом пространстве.

3. Интерпретация полученных осей (базовых характеристик) и конфигурации объектов - конечный результат применения многомерного шкалирования, дающий новое знание об изучаемой структуре (в случае корректного использования метода на всех этапах). Характер конфигурации объектов, а также внешние по отношению к исходным данным сведения позволяют дать содержательную интерпретацию осям и тем самым выявить глубинные мотивы, которыми руководствовались эксперты, упорядочивая пары объектов по степени их близости (в одном случае), или обнаружить скрытые факторы, определяющие структуру сходства и различия объектов (в другом случае).

Для методов многомерного шкалирования, как и для других методов анализа данных, слабо разработаны вероятностные модели и аппарат статистического оценивания.

Для повышения достоверности получаемых с помощью методов многомерного шкалирования результатов в одном исследовании нередко применяют различные методы многомерного шкалирования совместно с

другими методами; кластер-анализом, факторным анализом, множественной регрессией.

Задача многомерного шкалирования состоит в построении переменных на основе имеющихся расстояний между объектами. В частности, если нам даны расстояния между городами, программа многомерного шкалирования должна восстановить систему координат (с точностью до поворота и единицы длины) и приписать координаты каждому городу, так чтобы зрительно карта и изображение городов в этой системе координат совпали. Близость может определяться не только расстоянием в километрах, но и другими показателями, такими как размеры миграционных потоков между городами, интенсивность телефонных звонков, а также расстояниями в многомерном признаковом пространстве. В последнем случае задача построения такой системы координат близка к задаче, решаемой факторным анализом - сжатию данных, описанию их небольшим числом переменных. Нередко требуется, также, наглядное представление свойств объектов. В этом случае полезно придать координаты переменным, расположить в геометрическом пространстве переменные. С технической точки зрения это всего лишь транспонирование матрицы данных. Для определенности мы будем говорить о создании геометрического пространства для объектов, специально оговаривая случаи анализа множества свойств. В социальных исследованиях методом многомерного шкалирования создают зрительный образ «социально-экономического пространства» объектов наблюдения или свойств. Для такого образа наиболее приемлемо создание двумерного пространства.

Основная идея метода состоит в приписывании каждому объекту значений координат, так, чтобы матрица евклидовых расстояний между объектами в этих координатах, помноженная на константу оказалась близка к матрице расстояний между объектами, определенной из каких-либо соображений ранее.

Метод весьма трудоемкий и рассчитан анализ данных, имеющих небольшое число объектов.

Первая, в этом направлении, работа Торгерсона (Torgerson, 1952, [7]) была посвящена метрическому многомерному шкалированию. Модель этого метода имеет вид: $L\{S\}=D^2+E$ где $L\{S\}$ - линейное преобразование исходной матрицы расстояний, D^2 - матрица расстояний, полученная на основе созданных шкал, E - матрица отклонений модели от исходных данных. Линейное преобразование дает матрицу преобразованных расстояний $T=L\{S\}$. Целью многомерного метрического шкалирования является поиск оптимальных шкал и линейного преобразования матрицы исходных расстояний, минимизирующих ошибку Е.

Шепард и Краскэл (Shepard,1962, Kruscal,1964, [7]) совершили существенный прорыв, разработав метод неметрического шкалирования. Суть этого метода состоит в нелинейном преобразовании расстояний. Модель неметрического шкалирования имеет вид: $M\{S\}=D^2+E$, где $M\{S\}$ - монотонное преобразование исходной матрицы расстояний. Этот метод имеет больше шансов получить действительно геометрическое пространство, метрическое шкалирование. Монотонное преобразование дает матрицу преобразованных расстояний $T=L\{S\}$.

Для измерения качества подгонки модели Такейном (Takane, 1977) был предложен показатель $S\text{-}stress=\left(\frac{\|E\|}{\|T\|}\right)^{1/2}$, где норма матрицы $\|E\|$, $\|T\|$ означает сумму квадратов элементов матрицы. Слово stress в английском языке имеет множество значений, одно из этих значений - нагрузка. Этот показатель изменяется от 0 до 1. Равенство его нулю означает точную подгонку модели, единице - полную ее бессмысленность.

Кроме того, оценить качество модели можно с помощью показателя stress index Краскэла, который получается с использованием матрицы не квадратов расстояний, а расстояний.

Еще один показатель качества модели, RSQ, представляет собой квадрат коэффициента корреляции между матрицами T и E. Таким образом, также как в регрессионном анализе, RSQ может быть интерпретирован как доля дисперсии преобразованных расстояний T, объясненная матрицей расстояний D.

Можно построить для текущей конфигурации точек график зависимости воспроизведенных расстояния от исходных расстояний. Такая диаграмма рассеяния называется *диаграммой Шепарда*. По оси ординат OY показываются воспроизведенные расстояния (сходства), а по оси OX откладываются истинные сходства (расстояния) между объектами (отсюда обычно получается отрицательный наклон). На этом график также строится график ступенчатой функции. Ее линия представляет так называемые величины D-с крышечкой, то есть, результат монотонного преобразования исходных данных. Если бы все воспроизведенные результирующие расстояния легли на эту ступенчатую линию, то ранги наблюдаемых расстояний (сходств) был бы в точности воспроизведен полученным решением (пространственной моделью). Отклонения от этой линии показывают на ухудшение качества согласия (т.е. качества подгонки модели).

Чем больше размерность пространства, используемого для воспроизведения расстояний, тем лучше согласие воспроизведенной матрицы с исходной (меньше значение стресса). Если взять размерность пространства равной числу переменных, то возможно абсолютно точное воспроизведение исходной матрицы расстояний. Однако нашей целью является упрощение решаемой задачи, с тем, чтобы объяснить матрицу сходства (расстояний) в терминах лишь нескольких важнейших факторов (латентных переменных или вспомогательных шкал).

Тема 3. Многомерный регрессионный анализ.

Регрессионный анализ позволяет дать количественное описание взаимосвязей между переменными. Переменные, с которыми приходится работать являются случайными величинами. Случайные величины бывают:

1. дискретные
2. непрерывные, нормально распределенные
3. непрерывные, не подчиняющиеся нормальному распределению.

В связи с тем, что большинство величин хотя и дискретны, но имеют очень большое число возможных значений их считают непрерывными (доход населения, курсы валют, и т.д.). А наиболее часто встречается нормальное распределение непрерывных случайных величин, которое характеризуется тем, что большое скопление значений находится вблизи среднего значения, а при удалении от среднего количество наблюдений уменьшается. Нормальное распределение характеризуется кривой Гаусса.

Для оценки тесноты связи между двумя экономическими показателями в эконометрике используют коэффициент корреляции, т.к. при большом числе наблюдений почти все элементарные зависимости превращаются в линейные.

Поэтому наибольший интерес представляет линейная регрессия (зависимость).

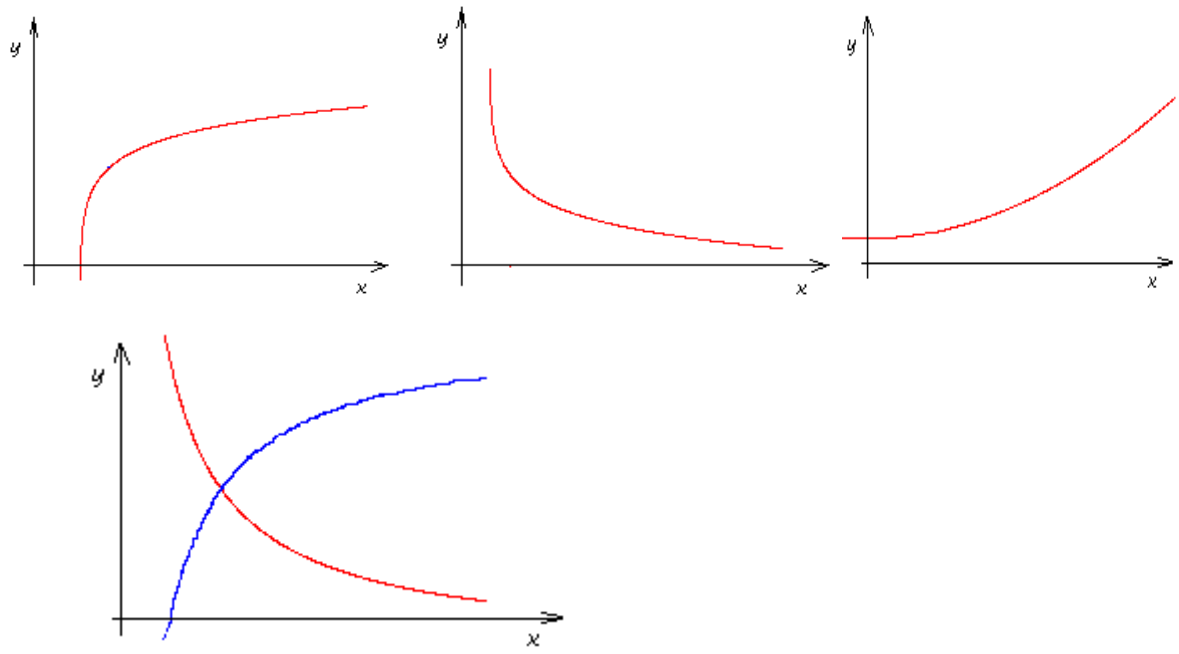
В зависимости от количества переменных (факторов), включенных в модель различают парную и множественную регрессии.

В силу многообразия и сложности социальных процессов многие зависимости не являются линейными и, следовательно, требуют моделирования нелинейными уравнениями.

Для случая парной регрессии подбор модели обычно осуществляется по виду расположения наблюдаемых точек на корреляционном поле, с учетом фактов известных из теории. В случае если зависимость может быть описана несколькими функциями, необходимо выбрать ту из них, которая обладает наилучшим качеством. Но следует помнить, что чем сложнее модель, тем менее интерпретируемы ее параметры.

Рассмотрим наиболее часто встречающиеся модели. Для простоты изложения и возможности графической интерпретации ограничимся моделями парной нелинейной регрессии.

При визуализации данных на корреляционном поле возможны следующие результаты:



1. Логарифмические модели.

– $\ln Y = \beta_0 + \beta \cdot \ln X + \varepsilon$ - *двойная логарифмическая модель*

коэффициент β в данной модели определяет эластичность переменной Y по переменной X, т.е. процентное изменение Y для данного процентного изменения X.

– $\ln y = a + b \cdot x + \varepsilon$ - *лог-линейная модель*, используется, например, при исследовании зависимости прироста объема выпуска от относительного увеличения затрат ресурса.

Коэффициент b в данной модели имеет смысл темпа прироста переменной y по переменной x, т.е. характеризует отношение относительного изменения y к абсолютному изменению x. Умножив b на 100%, получим процентный темп прироста переменной y.

– $y = a + b \cdot \ln x + \varepsilon$ - *линейно-логарифмическая модель*, используется, например, когда необходимо исследовать влияние процентного изменения независимой переменной на абсолютное изменение зависимой переменной.

Коэффициент b определяет изменение переменной y вследствии единичного относительного прироста x, например, если предположить, что y – валовой национальный продукт, а x – денежная масса, то b показывает, что с увеличением предложения денег на 1 % ВНП в среднем вырастет на b единиц.

2. Гиперболическая модель

$y = a + b \cdot \frac{1}{x} + \varepsilon$ - применяется в тех случаях, когда неограниченное увеличение значений объясняющей переменной x асимптотически приближает зависимую переменную y к некоторому пределу a. Подобная регрессия может отражать зависимости между объемом выпуска (x) и средними фиксированными издержками (y), между доходом (x) и спросом (y) на товары первой необходимости или предметы относительной роскоши (функция Торнквиста), между уровнем безработицы (x) и изменением заработной платы (y) и др.

3. Полиномиальная модель.

$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_k X^k$

При $k = 2$ квадратичная функция может отражать зависимость между расходами на рекламу (x) и прибылью (y), но большое применение имеет при анализе временных рядов; при $k = 3$ кубическая функция моделирует зависимость общих издержек (y) от объема выпуска (x).

4. Степенная модель.

$y = ax^b$ - отражает, например, зависимость спроса y на благо от его цены или от дохода x. Данная регрессия, путем математических преобразований сводится к двойной логарифмической модели. Коэффициент b является коэффициентом эластичности переменной y по переменной x.

5. Показательная модель.

$y = a \cdot e^{b \cdot x}$ - используется чаще всего в той ситуации, когда анализируется изменение переменной y с постоянным темпом прироста во времени.

Переменная x заменяется на t, а модель используется при анализе временных рядов.

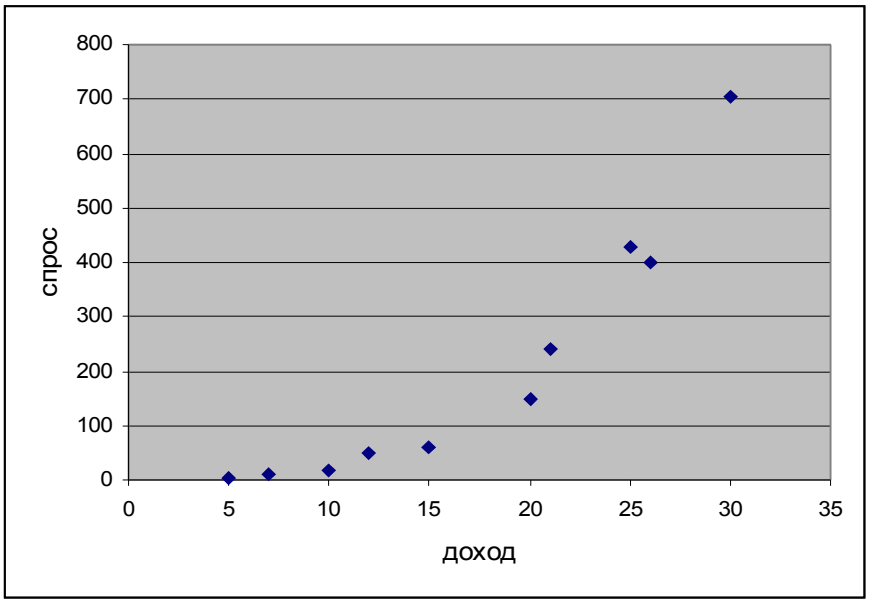
Коэффициент b показывает постоянный темп прироста переменной y во времени.

Показательная модель, путем логарифмирования сводится к лог-линейной модели.

Рассмотрим пример. Определить вид уравнения регрессии, описывающей зависимость между переменными оценить параметры регрессии, ее качество. Данные представлены в таблице

x - спрос	5	7	10	12	15	20	21	25	26	30
y - доход	3,51	9,5	27,4	47	91	213	245	412	462	705

Решение. Построим корреляционное поле, изобразив значения x и y в системе координат.



По корреляционному полю замечаем, что зависимость может быть описана функциями: $y = a \cdot e^{bx}$;

$y = ax^b$;

$y = ax^2 + bx + c$ и др.

Однако, по условию задачи переменная y – это спрос, а x – доход. Из теории известно, что зависимость между спросом на некоторое благо и доходом описывается функцией Энгеля вида $y = ax^b$. Поэтому уравнение регрессии для данных переменных y и x, будем искать в виде $y = ax^b$.

Линеаризируем функцию путем математических преобразований (логарифмирования).

$$y = ax^b$$

$$\ln y = \ln ax^b$$

$$\ln y = \ln a + \ln x^b$$

$$\ln y = \ln a + b \ln x$$

Обозначим $Y = \ln y$, $X = \ln x$, $A = \ln a$, получим

$$Y = A + bX$$

Оценим параметры полученного линейного уравнения методом наименьших квадратов. Необходимые расчёты представим в таблице:

№	x	y	X=ln x	Y=ln y	X^2	XY
1	5	3,51	1,61	1,26	2,59	2,02
2	7	9,5	1,95	2,25	3,79	4,38
3	10	17,8	2,30	2,88	5,30	6,63
4	12	50	2,48	3,91	6,17	9,72
5	15	60	2,71	4,09	7,33	11,09
6	20	150	3,00	5,01	8,97	15,01
7	21	240	3,04	5,48	9,27	16,69
8	25	430	3,22	6,06	10,36	19,52
9	26	400	3,26	5,99	10,62	19,52
10	30	705	3,40	6,56	11,57	22,31
сумма	171	2065,81	26,97	43,50	75,98	126,88
среднее	17,1	206,58	2,70	4,35	7,60	12,69

Для нахождения оценок параметров *A* и *b* составим систему уравнений:

$$\begin{cases} A + b \cdot \overline{X} = \overline{Y} \\ A \cdot \overline{X} + b \cdot \overline{X^2} = \overline{X \cdot Y} \end{cases} \quad \begin{cases} A + 2{,}7 \cdot b = 4{,}35 \\ 2{,}7 \cdot A + 7{,}6 \cdot b = 12{,}69 \end{cases}$$

решая систему уравнений получаем: *A=-3,61; b=2,95*

Зная, что $A = \ln a$ определим оценку параметра $a = e^A = e^{-3{,}61} = 0{,}027$, тогда уравнение зависимости примет вид: $y = 0{,}027 \cdot x^{2{,}95}$

Оценим качество полученной регрессии с помощью коэффициента детерминации:

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

	x	y	$\hat{y} = 0{,}027 \cdot x^{2{,}95}$	$(y - \hat{y})^2$	$(y - \bar{y})^2$
1	5	3,51	3,11	0,16	41237,83
2	7	9,5	8,40	1,20	38840,92
3	10	17,8	24,06	39,23	35638,27
4	12	50	41,20	77,35	24517,61
5	15	60	79,59	383,58	21485,99
6	20	150	185,95	1292,59	3201,41
7	21	240	214,74	638,12	1116,83
8	25	430	359,16	5018,44	49916,05
9	26	400	403,21	10,33	37410,91
10	30	705	614,99	8100,93	248421,50
сумма		2065,81		15561,95	501787,3
среднее		206,58			

$$R^2 = 1 - \frac{15561{,}95}{501787{,}3} = 0{,}967$$

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - k - 1}{k} = \frac{0{,}967}{1 - 0.967} \cdot \frac{10 - 1 - 1}{1} = 249{,}96$$

Значение коэффициента детерминации близко к 1, F-статистика указывает на значимость коэффициента детерминации. Полученная модель имеет отличное качество.

Коэффициент *b=2,95* в данной модели является коэффициентом эластичности, т.е. показывает, что с увеличением дохода потребителя на 1% спрос на товар увеличивается на *2,95* %.

Тема 4. Временные ряды.

Любой процесс можно описать с помощью уравнений. Наиболее распространенными являются уравнения линейной регрессии, временные уравнения и их системы.

Временные уравнения (ряды) определяют значения переменных в некоторый момент времени. В них возможно присутствие зависимой переменной в левой и правой части уравнения одновременно, но в разные моменты времени, модели данного типа называют динамическими.

Обычно динамические модели подразделяют на два класса

1. Модели с лагами (модели с распределенными лагами) – это модели, содержащие в качестве лаговых переменных лишь независимые (объясняющие) переменные

2. Авторегрессионные модели – это модели, уравнения которых в качестве лаговых объясняющих переменных включают значения зависимых переменных.

Временные ряды используются достаточно широко. Это вполне естественно, так как во многих случаях воздействие одних факторов на другие осуществляется не мгновенно, а с некоторым временным запаздыванием – лагом. Причин наличия лагов достаточно много, и среди них можно выделить следующие:

Психологические причины, которые обычно выражаются через инерцию в поведении людей. Например, люди тратят доход постепенно, а не мгновенно. Привычка к определенному образу жизни приводит к тому, что люди приобретают те же блага в течение некоторого времени даже после падения реального дохода.

Технологические причины. Например, изобретение персональных компьютеров не привело к мгновенному вытеснению ими больших ЭВМ в силу необходимости замены соответствующего программного обеспечения, которое потребовало продолжительного времени.

Институциональные причины. Например, контракты между фирмами, трудовые договоры требуют определенного постоянства в течение времени контракта (договора).

Механизмы, формирования показателей. Например, инфляция во многом является инерционным процессом; денежный мультипликатор (создание денег в банковой системе) также проявляет себя на определенном временном интервале и т.д.

Конечной целью анализа временных рядов является прогнозирование будущих значений исследуемого показателя. Такое прогнозирование позволяет, во-первых, предвидеть будущие реалии, во-вторых,

проанализировать построенную регрессионную модель на устойчивость (т.е. ее применимость в изменяющихся условиях).

Прогнозирование можно осуществлять либо на основе выявленных закономерностей изменения самого исследуемого показателя во времени и экстраполяции его прошлого поведения на будущее, либо на основе выявленной зависимости исследуемого показателя от других факторов, будущие значения которых контролируемы, известны или легко предсказуемы.

Различают долгосрочное и краткосрочное прогнозирование. В первом анализируется долговременная динамика исследуемого показателя, и в этом случае главным представляется выделение общего направления его изменения – тренда. При этом считается возможным пренебречь краткосрочными колебаниями значений показателя относительно этого тренда. Тренд обычно строится методами регрессионного анализа.

После выделения долгосрочного тренда обычно пытаются определить факторы, вызывающие отклонения значений исследуемой величины от тренда – колеблемость или сезонность. Для предсказания краткосрочных колебаний проводится более детальный регрессионный анализ с целью выявления большого числа показателей определяющих поведение исследуемой величины. Кроме этого, проводят более детальное исследование связей текущих значений исследуемых показателей с их прошлыми значениями или с прошлыми значениями других факторов.

Если амплитуда изменения переменной y с течением времени остается постоянной, можно предположить наличие аддитивной сезонности. В этом случае значение зависимой переменной в каждый момент времени найдем из условия: $y=T+S$. Если амплитуда изменения переменной y с течением времени увеличивается (уменьшается), можно предположить наличие мультипликативной сезонности. В этом случае значение зависимой переменной в каждый момент времени найдем из условия: $y=TS$, где T – тренд, S – сезонная компонента.

Мультипликативная сезонность Аддитивная сезонность

Для определения уравнения тренда используют регрессионный анализ, для оценки сезонной компоненты проводят центрирование данных.

Рассмотрим построение тренда с сезонностью на примере: поквартальные данные о прибыли некоторого предприятия в течение 6 лет представлены в таблице. Спрогнозировать прибыль предприятия на 2006 год поквартально.

Год	2000				2001				2002				2003				2004				2005			
Квартал	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
прибыль	11,86	14,34	15,22	14,53	10,84	13,78	16,41	15,68	12,19	15,65	17,69	16,37	13,21	16,4	17,89	16,85	14,08	16,3	18,31	17,7	14,73	17,74	19,44	18,23

Решение. Нам представлен временной ряд, т.к. одна переменная y - прибыль меняется с течением времени t.

По графику замечаем, что амплитуда изменения переменной y с течением времени остается постоянной, поэтому можно предположить наличие аддитивной сезонности.

В этом случае значение прибыли в каждый момент времени найдем из условия: y=T+S,

где T – тренд,

S – сезонная компонента.

Для оценки сезонной компоненты и определения уравнения тренда проведем центрирование данных, т.е. определим скользящую среднюю.

Вычислим суммарную прибыль за каждые 4 квартала путем поочередного суммирования значений переменной y, и среднее значение прибыли за год как $y_{ср} = \frac{y_{сум}}{4}$.

год	кварталы	y	итого за 4 квартала	среднее за 4 квартала	скользящая средняя	оценка сезонной компоненты
2000	1	11,86				
	2	14,34				
	3	15,22				
	4	14,53	55,95	13,99		
2001	1	10,84	54,93	13,73	13,86	-3,02
	2	13,78	54,37	13,59	13,66	0,12
	3	16,41	55,56	13,89	13,74	2,67
	4	15,68	56,71	14,18	14,03	1,65
2002	1	12,19	58,06	14,52	14,35	-2,16
	2	15,65	59,93	14,98	14,75	0,90
	3	17,69	61,21	15,30	15,14	2,55
	4	16,37	61,90	15,48	15,39	0,98
2003	1	13,21	62,92	15,73	15,60	-2,39
	2	16,40	63,67	15,92	15,82	0,58
	3	17,89	63,87	15,97	15,94	1,95
	4	16,85	64,35	16,09	16,03	0,82

Продолжение таблицы

2004	1	14,08	65,22	16,31	16,20	-2,12
	2	16,3	65,12	16,28	16,29	0,01
	3	18,31	65,54	16,39	16,33	1,98
	4	17,7	66,39	16,60	16,49	1,21
2005	1	14,73	67,04	16,76	16,68	-1,95
	2	17,74	68,48	17,12	16,94	0,80
	3	19,44	69,61	17,40	17,26	2,18
	4	18,23	70,14	17,54	17,47	0,76

Скользящая средняя определяется как среднее значение между двумя соседними значениями $y_{ср.}$ $y_{ск\,ср} = \frac{(y_{ср})_{i-1} + (y_{ср})_i}{2}$

Разница между соответствующими значениями прибыли и значениями скользящей средней определяет оценку сезонной компоненты.

Выпишем оценку сезонной компоненты в отдельную таблицу для расчета сезонности

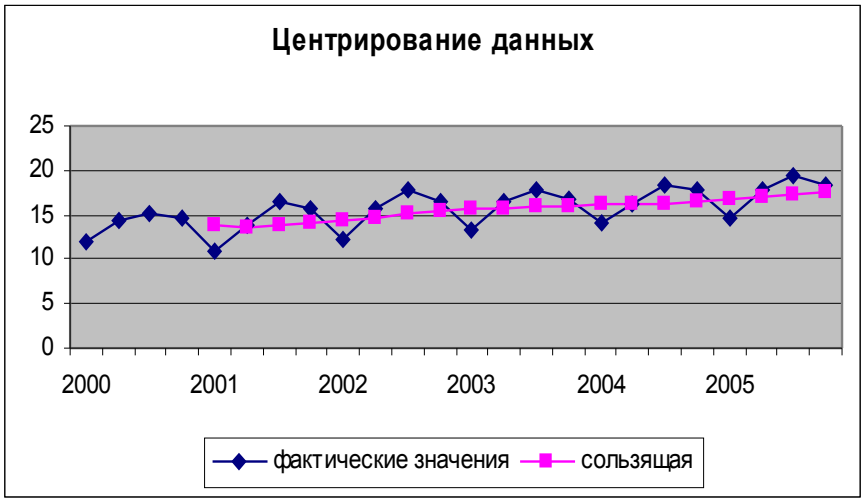
год	квартал			
	1	2	3	4
2000				
2001	-3,02	0,12	2,67	1,65
2002	-2,16	0,90	2,55	0,98
2003	-2,39	0,58	1,95	0,82
2004	-2,12	0,01	1,98	1,21
2005	-1,95	0,80	2,18	0,76
среднее	**-2,33**	**0,48**	**2,26**	**1,08**
среднее-$S_{ср}$	-2,70	0,11	1,89	0,71

Найдем среднюю оценку сезонной компоненты за год по формуле:

$$S_{ср} = \frac{\sum_{i=1}^{4} S_{ср\,i}}{4} = \frac{-2.33 + 0.48 + 2.26 + 1.08}{4} = 0.38$$

Для определения значений сезонной компоненты в каждом квартале необходимо среднюю оценку сезонной компоненты за год вычесть из соответствующих средних за каждый квартал. (последняя строка табл.)

Графическое изображение значений скользящей средней позволяет предположить наличие линейного тренда.



Уравнение линейного тренда временного ряда записывается в виде:

$y_t = \Delta y(абс) + y_{t-1}$ или $y_t = y_0 + \Delta y(абс) \cdot t$,где

$\Delta y(абс) = y_i - y_{i-1}$ - абсолютный прирост переменной y.

год	кварталы	y	$y_{ск}$	$\Delta y_{ск}$(абс)	$\Delta y_{ск}$(отн)	T	S	T+S
2000	1	11,86				13,10	-2,70	10,40
	2	14,34				13,29	0,11	13,40
	3	15,22				13,48	1,89	15,37
	4	14,53				13,67	0,71	14,38
2001	1	10,84	**13,86**			**13,86**	-2,70	11,16
	2	13,78	13,66	-0,20	0,99	14,05	0,11	14,15
	3	16,41	13,74	0,08	1,01	14,24	1,89	16,13
	4	15,68	14,03	0,29	1,02	14,43	0,71	15,14
2002	1	12,19	14,35	0,31	1,02	14,62	-2,70	11,92
	2	15,65	14,75	0,40	1,03	14,81	0,11	14,91
	3	17,69	15,14	0,39	1,03	15,00	1,89	16,89
	4	16,37	15,39	0,25	1,02	15,19	0,71	15,90
2003	1	13,21	15,60	0,21	1,01	15,38	-2,70	12,68
	2	16,40	15,82	0,22	1,01	15,57	0,11	15,67
	3	17,89	15,94	0,12	1,01	15,76	1,89	17,65
	4	16,85	16,03	0,08	1,01	15,95	0,71	16,66

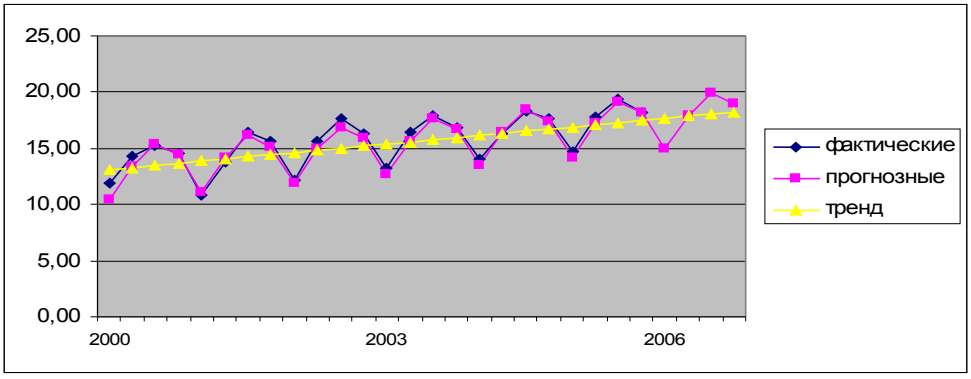
Продолжение таблицы

2004	1	14,08	16,20	0,17	1,01	16,14	-2,70	13,44
	2	16,30	16,29	0,10	1,01	16,33	0,11	16,43
	3	18,31	16,33	0,04	1,00	16,52	1,89	18,41
	4	17,70	16,49	0,16	1,01	16,71	0,71	17,42
2005	1	14,73	16,68	0,19	1,01	16,90	-2,70	14,20
	2	17,74	16,94	0,26	1,02	17,09	0,11	17,19
	3	19,44	17,26	0,32	1,02	17,28	1,89	19,17
	4	18,23	17,47	0,21	1,01	17,47	0,71	18,18
2006	*1*		среднее$\Delta y_{ск}$(абс)=**0,19**		**1,01**	*17,66*	*-2,70*	*14,96*
	2					*17,85*	*0,11*	*17,95*
	3					*18,04*	*1,89*	*19,93*
	4					*18,23*	*0,71*	*18,94*

Значения $\Delta y(отн) = \frac{y_i}{y_{i-1}}$ стечением времени практически не изменяются, что также указывает на наличие линейного тренда данного временного ряда.

В качестве $\Delta y(абс)$ выбираем среднее значение прироста переменной y равное 0,19. Тогда уравнение тренда имеет вид: $y_t = 0{,}19 + y_{t-1}$.Значения трендовой составляющей рассчитаны в таблице (столбец Т), путем прибавления к начальному значению скользящей средней последовательно по 0,19. Учитывая аддитивную сезонность и прибавляя значения сезонной компоненты (S) к каждому значению трендовой составляющей определяем прибыль предприятия и осуществляем прогноз на 4 квартала.

Графическое представление полученных данных позволяет судить о достаточно высоком качестве прогноза, т.к. отклонение прогнозной кривой от фактической несущественно.



Итак, при неизменных внешних условиях ожидаемая прибыль некоторого предприятия в 2006 году составит:

год	квартал	прибыль
2006	1	14,96
	2	17,95
	3	19,93
	4	18,94
итого		71,78

3.2. Задания для практических и домашних работ.

Дискриминантный анализ.

Имея достоверную информацию об обанкротившихся и не обанкротившихся в течении 5 лет предприятиях попытайтесь спрогнозировать поведение «молодых» фирм с помощью дискриминантного анализа.

Многомерное шкалирование

Задача. При устройстве на работу 18 претендентов прошли 10 различных тестов в кадровом отделе предприятия.

Номер теста	обозначение	Предмет теста
1	t1	Память на числа
2	t2	Математические задачи
3	t3	Находчивость при прямом диалоге
4	t4	Тест на составление алгоритмов
5	t5	Уверенность во время выступления
6	t6	Командный дух
7	t7	Находчивость
8	t8	Сотрудничество
9	t9	Признание в коллективе
10	t10	Сила убеждения

Максимальная оценка, которую можно было получить на каждом из тестов, составляет 10 балов. Результаты теста для 18 претендентов находятся в таблице в переменных t1-t10. Каждое наблюдение является характеристикой тестируемых кандидатов.

инициалы	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10
А.П.	10	9	3	10	4	5	1	7	6	5
М.К.	6	5	8	6	9	6	2	6	5	9
С.Т.	7	4	5	6	6	7	5	5	4	6
Т.Т.	8	8	5	6	6	4	6	7	7	5
А.А.	5	4	8	3	9	2	9	9	8	9
М.Р.	6	4	4	5	6	8	7	4	2	6
А.С.	9	10	7	8	6	9	3	6	5	7
Д.Н.	4	3	7	3	6	5	6	10	9	6
Ю.Т.	8	5	4	9	5	8	4	6	7	7
К.А.	8	6	4	6	5	7	8	6	8	5
Ч.Н.	7	7	6	7	7	9	9	3	4	7
М.Т.	5	5	9	5	10	5	10	6	5	9
Л.Г.	4	5	10	4	10	4	5	8	9	10
С.П.	10	10	6	9	5	8	4	8	10	5
Р.Д.	6	5	4	6	4	4	6	5	6	4
Ф.А.	6	7	3	7	2	3	8	4	3	3
К.С.	7	8	5	7	3	6	7	4	4	4
Г.Т.	8	9	6	8	5	10	7	10	8	5

С использованием результатов теста соответствия провести обнаружить группы кандидатов, близких по своим качествам.

Многомерный регрессионный анализ.

Задача 1 Имеются данные о стоимости однокомнатных квартир, где Price - цена, тыс. руб.; So - общая площадь, м^2; R - расстояние до центра, км

№	PRICE	SO	R
1	460	46	7
2	350	44	3,5
3	490	57,6	5
4	470	53,1	1,2
5	350	50	5
6	450	52,1	1,5
7	300	48	9
8	370	53	6,5
9	380	49	5
10	430	42	1,2
11	400	44	2
12	420	41	0

Оцените регрессии: $Price=\alpha+\beta*So$; $Price=\alpha+\beta*R$, $Price=\alpha+\beta*R+\gamma*So$ дайте объяснение полученным результатам.

Временные ряды. Сезонность.

Задача 1. Исследуется спрос населения *y* на прохладительные напитки в зависимости от дохода *x* по ежемесячным данным. Определить зависимость, отражающую увеличение спроса с ростом дохода. По фактическим точкам

наблюдений проверьте наличие сезонности. Добавьте переменную $s = \begin{cases} 0, & \textit{если холодное время года}, \\ 1, & \textit{если теплое время года}. \end{cases}$ и определите объем продаж с учетом сезона.

Задача 2. Небольшой частное кондитерское предприятие занимается производством печенья. В таблице представлены данные о динамике прибыли предприятия в течение нескольких лет. Построить трендовую модель. Спрогнозировать прибыль на 2003год.

№	*год*	*прибыль*
1	*1983*	*13,0*
2	*1984*	*13,2*
3	*1985*	*13,2*
4	*1986*	*14,2*
5	*1987*	*13,4*
6	*1988*	*13,8*
7	*1989*	*14,6*
8	*1990*	*14,2*
9	*1991*	*14,4*
10	*1992*	*16,6*
11	*1993*	*14,0*
12	*1994*	*15,4*
13	*1995*	*16,0*
14	*1996*	*16,8*
15	*1997*	*17,4*
16	*1998*	*18,1*
17	*1999*	*17,6*

3.3. Задания для самостоятельной работы студентов.

Студентам необходимо самостоятельно повторять ранее изученные понятия по математике и математическим методам в социологии из следующих разделов:

- линейная алгебра;
- линейное программирование;
- классические методы оптимизации;
- дисперсионный анализ;
- дифференциальные уравнения и их системы и др.

Кроме того, во время изучения дисциплины каждый студент самостоятельно выполняет **комплексное задание**: студенты составляют

задачу, связанную с их будущей профессиональной деятельностью и находят ее решение. Задача составляется на основании статистических данных, которые собирают студенты во время выполнения курсовых проектов по профессиональным дисциплинам. Таблица данных должна содержать 10-15 различных переменных, среди которых имеются фиктивные, и не менее 100 наблюдений. Средствами дискриминантного анализа и методами многомерного шкалирования все наблюдения разбиваются на классы по данным одной из фиктивных переменных (переменная указывается преподавателем). Для определенного класса строятся различные модели, и проводится их анализ. В результате выбирается та модель, которая наиболее полно отражает реальную картину.

Время, выделенное на самостоятельную работу, распределяется также на выполнение общих домашних заданий и подготовку к контрольным работам. Домашнее задание задается после каждого практического занятия и проверяется в начале следующего занятия.

3.4. Карта обеспеченности дисциплины кадрами профессорско-преподавательского состава.

шифр специальности	Обеспеченность преподавательским составом							
	Ф.И.О. должность по штатному расписанию	Какое образовательное учреждение профессионального образования окончил, специальность по диплому	Ученая степень и ученое звание (почетное звание)	Стаж научно педагогической работы			Основное место работы, должность	Условия привлечения к трудовой деятельности (штатный, совместитель (внутренний или внешний с указанием доли ставки), иное
				Всего	В т. ч. педагогический			
					Всего	В том числе по преподаваемой дисциплине		
1	2	3	4	5	6	7	8	9
040201	Двоерядкина Н.Н., доцент	БГПУ, учитель математики	к.п.н.	9	9	5	АмГУ	1 ставка