Федеральное государственное автономное образовательное учреждение высшего образования
"Национальный исследовательский университет "Высшая школа экономики"

Факультет социальных наук

Программа дисциплины «Многомерный статистический анализ»

для направления 41.03.04 «Политология» подготовки бакалавра

Авторы программы:
Камалова Р. У., преподаватель (rkamalova@hse.ru)
Сальникова Д.В., преподаватель (dsalnikova@hse.ru)

Одобрена на заседании кафедры высшей математики «___»_____ 2017 г.
Зав. кафедрой к.ф.-м.н., проф. Макаров А.А.

Рекомендована секцией УМС _____«___»_____ 2017 г.
Председатель _____

Утверждена УС факультета прикладной политологии «___»_____2017 г.
Ученый секретарь _____

Москва, 2017

1 Область применения и нормативные ссылки

Настоящая программа учебной дисциплины устанавливает минимальные требования к знаниям и умениям студента и определяет содержание и виды учебных занятий и отчетности.

Программа предназначена для преподавателей, ведущих данную дисциплину, учебных ассистентов и студентов направления 41.03.04 «Политология» подготовки бакалавра, изучающих дисциплину «Многомерный статистический анализ».

2 Цели освоения дисциплины

Цель освоения дисциплины «Многомерный статистический анализ» – выработать базовые компетенции по решению задач, связанных с анализом эмпирических данных с помощью методов многомерной статистики.

В соответствии с поставленной целью, курс решает следующие задачи:

а) формирование у студентов знания понятий и идей, лежащих в основе многомерной математической статистики;

б) освоение основных статистических моделей социально-экономических и политических процессов и явлений;

в) овладение основными методами многомерной математической статистики, позволяющими решать различные социально-экономические и политологические исследовательские задачи;

г) формирование у студентов понимания перспектив использования статистических методов анализа данных в прикладной политологии.

3 Компетенции обучающегося, формируемые в результате освоения дисциплины

В результате освоения дисциплины студент должен:

а) владеть методами многомерной статистики в объеме данной программы;

б) уметь применять изученные в рамках дисциплины методы многомерной статистики к решению содержательных социально-экономических и политологических задач в соответствующем программном обеспечении и содержательно интерпретировать полученные результаты;

в) также знать обязательную литературу в полном объеме.

Компетенция	**Код по ФГОС / НИУ**	**Дескрипторы – основные признаки освоения (показатели достижения результата)**	**Формы и методы обучения, способствующие формированию и развитию компетенций**
Свободное владение литературной и деловой письменной и устной речью на русском языке, навыками публичной и научной речи; умение создавать и редактировать тексты профессионального назначения, анализировать логику рассуждений и высказываний	ОК-2	Демонстрирует способность к критическому анализу исследований, представлению результатов самостоятельного анализа данных в виде статьи	Текущие домашние задания: чтение и критический анализ научных статей с последующим обсуждением на семинарах Домашнее задание: самостоятельное исследование с использованием изученных методов

Владение иностранным языком на уровне, достаточном для профессионального общения; для поиска и анализа иностранных источников информации	ОК-3	Владеет английским языком на уровне, достаточном для чтения научных статей в зарубежных журналах	Текущие домашние задания: чтение и критический анализ научных статей с последующим обсуждением на семинарах Домашнее задание: подготовка обзора литературы для проведения самостоятельного исследования
Понимание основных положений и методов социальных, гуманитарных и экономических наук, способность использовать их при решении социальных и профессиональных задач, способность анализировать значимые социальные и экономические проблемы и процессы	ОК-11	Применяет полученные знания из области социальных, гуманитарных и экономических наук для критического анализа статей, постановки исследовательской задачи и последующей интерпретации полученных результатов исследования	Домашнее задание: постановка исследовательской задачи, интерпретация полученных результатов исследования
Способность применять методы математического анализа и моделирования для решения задач профессиональной деятельности	ОК-12	Применяет изученные в рамках курса методы многомерного статистического анализа, интерпретирует полученные результаты	Семинарские занятия и домашние задания
Способность к участию в научных исследованиях политических процессов и отношений, владение методами анализа и интерпретации представлений о политических явлениях на различных уровнях организации мира	ПК-1	Применяет полученные знания о политических явлениях при проведении самостоятельного исследования	Эссе: проведение самостоятельного исследования

4. Место дисциплины в структуре образовательной программы

Для указанного направления подготовки дисциплина является обязательной.
Изучение данной дисциплины базируется на следующих дисциплинах:

- «Теория вероятностей и математическая статистика» (1 курс)

- «Вероятностно-статистические модели в политологии» (2 курс),

Основные положения дисциплины должны быть использованы в дальнейшем при изучении следующих дисциплин:

- «Политический анализ»
- «Политическая регионалистика»
- «Оценка эффективности мер социальной политики государства»
- «Коллективный выбор: теория и эмпирические исследования»

5. Тематический план учебной дисциплины

№	Наименование разделов	Всего часов	Аудиторные часы		Самостоятельная работа
			Лекции	Семинары	
1	Снижение размерности признакового пространства	80	12	18	50
2	Методы классификации	58	8	10	40
3	Анализ панельных данных	90	12	18	60
	ИТОГО	**228**	**32**	**46**	**150**

6 Формы контроля знаний студентов

Тип контроля	**Форма контроля**	**1 год**				**Параметры**
		1	**2**	**3**	**4**	
Текущий	Контрольная работа	*				Письменная работа 80 минут
	Эссе		*			Выполненное самостоятельно студентом исследование посредством изученных методов многомерного статистического анализа

						Объем домашнего задания: 20000 – 40000 знаков с пробелами.
Итоговый	Экзамен		*			Письменная работа (160 минут)

6.1 Критерии оценки знаний, навыков

Оценки за домашние задания и зачет выставляются, исходя из следующих критериев:

- правильность решения задачи (корректный выбор метода/методов),
- полнота решения задачи (понимание допущений для реализации метода/методов, оценка качества полученных результатов),
- наличие и корректность интерпретации полученных результатов.

Оценки по всем формам текущего и итогового контроля выставляются по 10-ти балльной шкале.

Письменный экзамен состоит из двух частей: письменной и работы за компьютером. В письменной части будут представлены задачи, проверяющие базовые навыки и элементарные компетенции. Для работы за компьютером будут представлены задания, предполагающие проверку способностей и применять к решению содержательных задач методы многомерного статистического анализа (используя R), а также формулировать содержательные выводы (в том числе в форме кратких ответов на вопросы преподавателя).

7 Содержание дисциплины

Тема 1. Снижение размерности многомерного признакового пространства (построение индексов)

Индекс как результат снижения размерности многомерного признакового пространства. Объяснение потребности в индексах. Размерность каких признаковых пространств может быть снижена? Постановка задачи метода главных компонент (МГК). Ковариационная / корреляционная матрица как основной объект. Алгоритм МГК.

Главная компонента vs. среднее. Свойства главных компонент (взаимная ортогональность; наименьшее искажение геометрической структуры данных, наименьшая ошибка автопрогноза). Оценка качества снижения размерности: доля объясненной вариации и понятие информативность главной компоненты, методы Г. Кайзера и Р.Б. Кеттелла.

Факторный анализ. Разведывательный и конфирматорный факторный анализ.

Введение в моделирование структурными уравнениями.

Основная литература:

Principal Component Analysis. In *Analysis of Multivariate Social Science Data* edited by David J. Batholomew, Fiona Steele, Irini Moustaki and Jane I. Galbraith. Boca Raton, London, New York: CRC Press, 2008. Pp. 117 – 144.

Jolliffe I.T. 2002. Principal Component Analysis. New York: Springer. Chapters 1, 6.

Айвазян С.А., Мхитарян В.С. Прикладная статистика. Основы эконометрики. Т.1: Теория вероятностей и прикладная статистика. - М.: ЮНИТИ, 2001. – С. 520 – 550.

Дополнительная литература

Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы. – М.: Финансы и статистика, 2003. – 352 с.

Тема 2. Методы классификации

Классификация как одна из общенаучных задач. Типология задач классификации: с обучением и без обучения, параметрическая и непараметрическая постановка задачи.
Постановка задачи иерархической кластеризации. Понятия расстояния и его свойства. Виды расстояний и проблема их выбора. Понятие типа (алгоритма) агломерации и его виды. Проблема выбора алгоритма агломерации. Определение числа кластеров в задаче иерархического кластерного анализа (дендрограмма и сосульчатая диаграмма *(icicle plot)*). Проблема устойчивости результатов. Применение кластерного анализа к классификации поведения объектов в динамике. Методы валидизации результатов кластеризации: дисперсионный анализ, лепестковые диаграммы.
Расщепление смесей вероятностных распределений – параметрический метод без обучения. Условия применимости, оценивание параметров распределений и решение о групповой принадлежности единиц наблюдения.
Дискриминантный анализ – параметрический метод с обучающей выборкой. Дискриминантные функции. Линейный дискриминантный анализ Фишера. Проверка допущений о равенстве ковариационных матриц. Интерпретация коэффициентов дискриминантных функций. Сравнение с моделями дискретного выбора.

Основная литература:

Cluster Analysis. In *Analysis of Multivariate Social Science Data* edited by David J. Batholomew, Fiona Steele, Irini Moustaki and Jane I. Galbraith. Boca Raton, London, New York: CRC Press, 2008. Pp. 17 – 53.

Gore Paul A., Jr. Cluster Analysis. In *Handbook of Applied Multivariate Statistics and Mathematical Modeling*. Academic Press, 2000. Pp. 297 – 321.

Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. Т.1: Теория вероятностей и прикладная статистика. – М.: ЮНИТИ, 2001 – С. 471 – 478, 479 – 488, 488 – 518.

Centellas, Miguel and Mihaiela Ristei Gugiu. (2013). The Democracy Cluster Classification Index. Political Analysis, 21: 334–349.

Дополнительная литература

Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы. – М.: Финансы и статистика, 2003. – С. 241 – 254.

Ким О. Дж., и др. Факторный, дискриминантный и кластерный анализ. — М.: Финансы и статистика, 1989. — с. 78-138, 139-209.

Rui Xu, Don Wunsch. Clustering. Wiley-IEEE Press, 2009. Pp. 1-110.

Tabachnick, B. G., and Fidell, L. S. Using Multivariate Statistics, 6th ed. Boston: Allyn and Bacon. 2013, Pp. 377-438.

Тема 3. Анализ панельных данных

Панельная структура данных VS пространственно-временные данные (time-series cross-section data). Модель с фиксированными эффектами. Внутригрупповое преобразование. Модель со случайными эффектами. Выбор адекватной модели: F-тест, тест множителей Лагранжа Бреуша-Пагана, тест Хаусмана.

Обобщенный метод наименьших квадратов (GLS). Реализуемый обобщенный метод наименьших квадратов. Панельно-скорректированные стандартные ошибки.

Пространственная корреляция: суть, условия возникновения, последствия. Тест Pesaran. Тест Frees.

Временная автокорреляция. Последствия автокорреляции. Диагностика: статистика Дарбина-Уотсона, тест Бреуша-Годфри. Подход к автокорреляции в рамках классической эконометрической школы (serial correlation as a nuisance). Процедура Кохрейна-Оркатта, процедура Прайс-Уинстена. Стандартные ошибки Ньюи-Уэста.
Динамические модели. Функция импульсного отклика. Модели с включением лагированных независимых переменных. Модели с включением лагированного отклика в качестве предиктора. Процедура Ареллано – Бонда. Обобщенный метод моментов.

Основная литература

Allison P. 2009. Fixed Effects Regression Models. Thousand Oaks, CA: Sage Publications. Pp. 1 – 27.

Beck N., Katz J. N. (1995). What to Do (and Not to Do) with Time-Series-Cross-Section Data in Comparative Politics. American Political Science Review, Vol. 89, Issue 3, pp. 634 – 647.

Beck N., Katz J. N. (2011). Modeling Dynamics in Time-Series-Cross-Section Political Economy Data. Annual Review of Political Science, Vol. 14. Pp. 331 – 352.

Gujarati, D.N. Basic econometrics. New York McGraw-Hill, 2003

Дополнительная литература

Вербик М. Путеводитель по современной эконометрике. Пер. с англ. В. А. Банникова. Под науч. ред. и предисл. С.А. Айвазяна. – М.: Научная книга, 2008.

Ратникова Т.А., Фурманов К.К. Анализ панельных данных и данных о длительности состояний. Уч. Пособие. – М.: изд. Дом Высшей школы экономики, 2014.

Boef de S., Keele L. (2008). Taking Time Seriously. American Journal of Political Science, Vol. 52, No. 1, pp. 184 – 200.

Green D. P., Kim S.Y., Yoon D.H. (2001). Dirty Pool. International Organization. Vol. 55, No.2, pp. 441 – 468.

Hoyos R., Sarafidis V. (2006). Testing for Cross-Sectional Dependence in Panel-Data Models. The Stata Journal. Vol. 6, No. 4. Pp. 482 – 496.

8 Образовательные технологии

Занятия проводятся в форме лекций и практических занятий в компьютерном классе.

9 Оценочные средства для текущего контроля и аттестации студента

9.1 Тематика заданий текущего контроля

Домашняя работа

Домашнее задание – самостоятельная (индивидуальная) письменная работа по результатам поиска и изучения соответствующей литературы и анализа массива политологических и социально-экономических данных с использованием изученных методов многомерной статистики.

Рекомендуемый объем домашнего задания – 15-35 тыс. символов с пробелами.

Структура домашнего задания в обязательном порядке должна включать в себя следующие разделы:

1. Аннотация (от 500 до 800 знаков с пробелами)
2. Введение (постановка задачи, обзор литературы и краткая «приманка» для читателя в виде формулировки основных полученных в работе результатов)
3. Гипотезы (формулируются и обосновываются гипотезы работы)
4. Данные и используемые методы (описываются имеющиеся данные и используемые методы; выбор методов обосновывается)
5. Результаты (в табличной форме, соответствующей зарубежным публикационным стандартам, представляются полученные результаты, а также предлагается их интерпретация)
6. Обсуждение полученных результатов (описывается, в какой мере полученные результаты соответствуют существующей литературе; обсуждаются возможные причины расхождений; обсуждается устойчивость полученных результатов к изменению массива данных и изменению методов анализа)
7. Заключение (формулируются выводы работы, обсуждаются теоретические следствия из полученных выводов, указываются возможные направления дальнейшего уточнения результатов)
8. Список использованной литературы

Оцениваются адекватность формулировки задачи/проблемы, соответствие заявленных для проверки гипотез сформулированной задаче, корректность подбора методов и интерпретации результатов, обоснованность выводов. Обращайте внимание на грамотность русского языка и корректность использования терминов, аккуратность оформления, в т.ч. библиографии.

Если объем и характер заимствования, оформленного в виде ссылки, ставят под сомнение самостоятельность выполнения работы, преподаватель действует в соответствии с п. 2.5 Положения о плагиате НИУ ВШЭ.

Задание выполняется в R.

Оценка выставляется по 10-балльной шкале.

К работе также необходимо приложить массив данных, R-script.

9.2 Примерные вопросы для оценки качества освоения дисциплины

1. В чем заключается проблема смещения самоотбора при анализе панельных данных? Проиллюстрируйте свой ответ примером.
2. Назовите допущения модели со случайными эффектами (как допущения об индивидуальных эффектах u_i, так и о следующей составляющей случайной ошибки e_{it}).
3. Обозначьте недостатки оценивания модели посредством РОМНК (FGLS).
4. Каковы последствия гетероскедастичности без коррекции стандартных ошибок или реализации ОМНК (GLS)?
5. Для чего используются панельно-скорректированные стандартные ошибки: какие проблемы они позволяют решить?
6. В чем заключается проблема пространственной корреляции?
7. Опишите способы диагностики автокорреляции.
8. Почему нельзя включать лаги отклика в модель с фиксированными эффектами в условиях недостаточного количества временных периодов? (иными словами, к чему это приводит?)
9. Покажите посредством построения соответствующей функции импульсного отклика, что включение лага отклика (y) в модель означает, что эффект независимой переменной (x) бесконечен.
10. Для решения какой задачи применяется кластерный анализ?
11. Сформулируйте свойства, которым должно удовлетворять любое расстояние. Какое из этих свойств выполняется не всегда (например, в психологических исследованиях)?
12. Какие виды метрики (расстояний) Вам известны?
13. Какую метрику следует использовать при кластеризации количественных признаков методом Варда (Ward)?
14. Какую метрику Вы бы предложили использовать для кластеризации выборки женщин по следующим переменным: курение табака (да, нет), семейный статус (не замужем, замужем/проживаем совместно, разведена, вдова), количество детей, должность (руководящая, не руководящая)? Объясните почему. Напишите формулу для расчета.
15. Почему не очень осмысленно применять алгоритмы кластерного анализа для классификации объектов в одномерном или двумерном признаковом пространстве?
16. Назовите основные требования, необходимые для реализации процедуры расщепления смеси вероятностных распределений.
17. Назовите задачи, которые решает метод главных компонент (МГК).
18. Дайте определение главной компоненте.
19. Чему равен коэффициент корреляции между главными компонентами?
20. Для какого из наборов данных процедура МГК, основанная на ковариационной матрице исходных признаков, не может быть применена по техническим причинам?
    - возраст; размер заработной платы; среднее количество часов, проверенное в интернете в день
    - размер заработной платы; общее количество лет обучения, включая школу; количество рабочих часов в неделю
    - размер заработной платы; общее количество лет обучения, включая школу; количество рабочих часов в неделю; характер занятости (самозанятость, работа по найму, госслужба)
    - доля затрат на образование; размер ВВП/чел.; уровень младенческой смертности
    - уровень безработицы; размер ВВП/чел.; размер дефицита гос. бюджета.

21. Опишите методы Г. Кайзера (1974 г.) и Р.Б. Кеттелла (1966 г.) определения числа главных компонент извлекаемых из набора признаков, подлежащих анализу методом главных компонент.
22. Чем отличаются конфирматорный и разведывательный факторный анализ?
23. Чем отличается факторный анализ от метода главных компонент?

10 Порядок формирования оценок по дисциплине

Преподаватель оценивает работу студентов на практических занятиях: активность в дискуссиях, правильность решения задач на семинаре, правильность и своевременность решения задач в текущих домашних заданиях и прочих заданиях, которые выдаются на семинарских занятиях. Оценки за работу на практических занятиях преподаватель выставляет в рабочую ведомость.

Накопленная оценка по 10-ти балльной шкале за работу на семинарских и практических занятиях определяется перед итоговым контролем. Накопленная оценка учитывает результаты студента по текущему контролю следующим образом:

$$Онакопл. = кк/р \cdot Ок/р + кэссе \cdot Од/з + кауд \cdot Оауд.,$$

где $кк/р = 0.2$, $кэссе = 0.5$, $кауд = 0.3$.

Округление каждого компонента накопленной оценки производится в соответствии с правилами математики и происходит до расчета накопленной оценки. Округление накопленной оценки также производится в соответствии с правилами математики.

Результирующая оценка за дисциплину рассчитывается следующим образом:

$$Орезульт = Онакопл. + Оэкзамен,$$

где $кнакопл. = 0.5$, $кэкзамен = 0.5$.

Округление результирующей оценки также производится в соответствии с правилами математики.

В диплом выставляется результирующая оценка по учебной дисциплине.

11 Учебно-методическое и информационное обеспечение дисциплины

11.1 Базовые учебники

Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. Т.1: Теория вероятностей и прикладная статистика. – М.: ЮНИТИ, 2001.

Allison P. 2009. Fixed Effects Regression Models. Thousand Oaks, CA: Sage Publications. Pp. 1 – 27.

Analysis of Multivariate Social Science Data, edited by David J. Batholomew, Fiona Steele, Irini Moustaki and Jane I. Galbraith. Boca Raton, London, New York: CRC Press, 2008.

Beck N., Katz J. N. (1995). What to Do (and Not to Do) with Time-Series-Cross-Section Data in Comparative Politics. American Political Science Review, Vol. 89, Issue 3, pp. 634 – 647.

Beck N., Katz J. N. (2011). Modeling Dynamics in Time-Series-Cross-Section Political Economy Data. Annual Review of Political Science, Vol. 14. Pp. 331 – 352.

Centellas, Miguel and Mihaiela Ristei Gugiu. (2013). The Democracy Cluster Classification Index. Political Analysis, 21: 334–349.

Gore Paul A., Jr. Cluster Analysis. In *Handbook of Applied Multivariate Statistics and Mathematical Modeling*. Academic Press, 2000. Pp. 297 – 321.

Gujarati, D.N. Basic econometrics. New York McGraw-Hill, 2003

Jolliffe I.T. 2002. Principal Component Analysis. New York: Springer.

11.2 Дополнительная литература

Вербик М. Путеводитель по современной эконометрике. Пер. с англ. В. А. Банникова. Под науч. ред. и предисл. С.А. Айвазяна. – М.: Научная книга, 2008.

Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы. – М.: Финансы и статистика, 2003. – С. 241 – 254.

Ким О. Дж., и др. Факторный, дискриминантный и кластерный анализ. — М.: Финансы и статистика, 1989. — с. 78-138, 139-209.

Ратникова Т.А., Фурманов К.К. Анализ панельных данных и данных о длительности состояний. Уч. Пособие. – М.: изд. Дом Высшей школы экономики, 2014.

Boef de S., Keele L. (2008). Taking Time Seriously. American Journal of Political Science, Vol. 52, No. 1, pp. 184 – 200.

Green D. P., Kim S.Y., Yoon D.H. (2001). Dirty Pool. International Organization. Vol. 55, No.2, pp. 441 – 468.

Hoyos R., Sarafidis V. (2006). Testing for Cross-Sectional Dependence in Panel-Data Models. The Stata Journal. Vol. 6, No. 4. Pp. 482 – 496.

Rui Xu, Don Wunsch. Clustering. Wiley-IEEE Press, 2009. Pp. 1-110.

Tabachnick, B. G., and Fidell, L. S. Using Multivariate Statistics, 6th ed. Boston: Allyn and Bacon. 2013, Pp. 377-438.

11.3 Программные средства

Для успешного освоения дисциплины, студент использует программное обеспечение R.

11.4 Дистанционная поддержка дисциплины

При выполнении домашних работ студентам рекомендуется пользоваться материалами

- Единого архива экономических и социологических данных НИУ ВШЭ (http://sophist.hse.ru/),
- Межуниверситетского консорциума по политическим и социальным исследованиям (ICPSR) (http://www.icpsr.umich.edu/icpsrweb/ICPSR/index.jsp),
- Проекта «Разработка учебно-методических материалов для преподавания курсов по применению количественного инструментария к решению социально-экономических задач» (http://www.hse.ru/jesda/mathbase/).

- World Development Indicators (World Bank) (http://library.hse.ru/e-resources/e-resources.html).
- World Values Survey (http://www.worldvaluessurvey.org/).
- Росстат (http://www.gks.ru/).
- Resources to help you learn and use R. UCLA: Academic Technology Services, Statistical Consulting Group. (https://stats.idre.ucla.edu/other/dae/).

12 Материально-техническое обеспечение дисциплины

Программное обеспечение для анализа данных: R. Возможно самостоятельное использование других программ: Stata, Python.