КАЗАНСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

Е.И. КАДОЧНИКОВА, Ю.А. ВАРЛАМОВА

СТАТИСТИЧЕСКИЙ АНАЛИЗ ПРОСТРАНСТВЕННЫХ ДАННЫХ

Учебное пособие

КАЗАНЬ
2023

УДК 330.43(075.8)
ББК 65.051я73
К13

Печатается по рекомендации учебно-методической комиссии Института управления, экономики и финансов Казанского (Приволжского) федерального университета (протокол № 7 от 27 февраля 2023 г.)

Рецензенты:
доктор технических наук, профессор **И.И. Исмагилов**;
кандидат экономических наук, доцент **Е.В. Козоногова**

Кадочникова Е.И.
К13 Статистический анализ пространственных данных: учебное пособие / Е.И. Кадочникова, Ю.А. Варламова. – Казань: Издательство Казанского университета, 2023. – 140 с.

ISBN 978-5-00130-700-6

Данное учебное пособие составлено в соответствии с современной структурой изучения аналитических, картографических и географических дисциплин, является базовым для курса «Статистический анализ пространственных данных».

Учебное пособие представляет собой первую, теоретическую часть учебного курса, в нем разъяснены основные термины и приемы методов статистического анализа пространственных данных, представлены базовые понятия пространственной эконометрики.

Учебное пособие предназначено для студентов вузов, аспирантов и преподавателей.

УДК 330.43(075.8)
ББК 65.051я73

ISBN 978-5-00130-700-6

Содержание

ТЕМА 1. ОСНОВНЫЕ ПОНЯТИЯ ПРОСТРАНСТВЕННОГО АНАЛИЗА

В результате освоения темы обучающийся будет:

– ***знать*** сущность, задачи и этапы проведения пространственного анализа;

– ***уметь*** определять тип данных под конкретные задачи исследования;

– ***владеть*** навыками составления плана проведения пространственного анализа в соответствии с выделенными этапами.

Основные вопросы:

1.1. Пространственный анализ: сущность, задачи, этапы.

1.2. Типы данных.

Ключевые слова: пространственный анализ, данные, количественные данные, качественные данные, географически структурированные данные, пространственно организованные данные

1.1. Пространственный анализ: сущность, задачи, этапы

Пространственный анализ – направление географических исследований, направленное на анализ свойств объекта, продиктованных его взаимосвязанностью с другими объектами. Пространственный анализ тесно взаимосвязан с пространственной статистикой и пространственной эконометрикой, активно использует методы компьютерного моделирования в геоинформационных системах (ГИС) с пространственной эконометрикой и математической статистикой.

Философия пространственного анализа базируется на понимании того, что свойства изучаемых объектов связаны с положением других объектов в пространстве – горизонтальная обусловленность. Горизонтальная обусловленность объекта находится в неразрывной связи с вертикальной обусловленностью – положение объекта в про-

странстве и его свойства, определяемые этим положением[1]. В качестве примера можно привести региональную экономику отдельного субъекта Российской Федерации, например, Республики Татарстан, уровень развития которой определяется не только природными ископаемыми на территории региона, но и выгодным экономико-географическим положением (исторически торговые пути проходили по территории Республики Татарстан, в том числе Шелковый путь).

Следует отметить, что существуют как сторонники, так и противники пространственного анализа. Можно выделить две философские концепции, которые находятся как бы на противоположных полюсах отношения к пространственным исследованиям. ***Географический детерминизм*** исходит из того, что расположение в пространстве является ведущим (иногда даже единственным) фактором, объясняющим свойства объекта. Согласно ***географическому нигилизму*** географический фактор не влияет на свойства объекта. Согласимся с мнением И.Ю. Окунева о том, что «истина находится примерно посередине, в ***географическом поссибилизме*** – пространство задает вероятностные сценарии для формирования свойства объекта, но не может однозначно определить его природу»[2].

Сущность пространственного анализа основывается на предположении, что на свойства, характеристики конкретного объекта или явления оказывает влияние аналогичные свойства и характеристики пространственно-взаимосвязанных объектов или явлений (или других явлений). Например, мы можем выдвинуть гипотезу, что уровень занятости населения в данном географическом районе зависит как от ситуации на рынке труда в данном районе (заработная плата, наличие вакансий, численность рабочей силы), так и от ситуации на локальных рынках труда в соседних районах. Пространственная взаимосвязь районов может поддерживаться транспортным сообщением между районами, культурно-исторической взаимосвязанностью (менталитет, язык, традиции), наличием финансовых потоков (донор-

[1] *Окунев И.Ю.* Основы пространственного анализа. М., 2020. С. 4.
[2] *Окунев И.Ю.* Основы пространственного анализа. М., 2020. С. 9.

реципиент), что создает возможности для миграции населения. Соответственно, ***задачами*** пространственного анализа являются:

1) выявление закономерностей в пространственных данных;
2) выделение однородных и неоднородных групп объектов;
3) определение пространственной взаимосвязи в данных;
4) моделирование пространственных причинно-следственных связей.

Следовательно, в пространственном анализе можно выделить следующие этапы, которые в зависимости от решаемой исследовательской задачи, реализуются полностью или частично (см. рис. 1.1).

Рис. 1.1. Этапы пространственного анализа

Пространственный анализ позволяет решать сложные локационно-ориентированные задачи, находить закономерности, оценивать тенденции и принимать решения, изучать характеристики различных местоположений и существующие взаимосвязи, также он добавляет новые возможности для принятия оптимальных решений.

С помощью пространственного анализа возможно сделать выводы о пространственных закономерностях распределения различных показателей, выявить ключевые факторы, влияющие на значения переменных.

Пространственный анализ применяет методы математической статистики и пространственной эконометрики. Так, на этапе подготовки данных проводится геокодирование данных, их привязка к координатам в пространстве. Кроме того, на данном этапе применяются:

1) математические преобразования (логарифмирование, взятие разностей для временных рядов);

2) нормирование – приведение к одному масштабу;
3) корректировка с учетом сезонности;
4) корректировка с учетом инфляции;
5) календарная корректировка.

Основная цель данного этапа пространственного анализа – выявить особенности используемых данных и привести их к виду, позволяющему нивелировать побочные эффекты.

На этапе визуализации данных в пространственном анализе часто используют построение различного рода картограмм, графиков, диаграмм. При работе с географическими данными можно использовать специальные пакеты QGIS, ArcGis и GeoDa.

Группировка данных позволяет типологизировать данные по определенным критериям. Так, например, группировка регионов России по географическому признаку позволяет выделить «западные» и «восточные» регионы, по уровню экономического развития – с высоким, средним и низким уровнем.

Пространственный анализ, который осуществляется не по одному свойству объекта, а по нескольким является многомерным и предполагает нахождение пространственной взаимосвязи между исследуемыми свойствами. Влияет ли уровень занятости в соседних регионах на аналогичный показатель в данном регионе? Ответы на подобные вопросы можно найти с помощью расчетов специальных коэффициентов пространственной автокорреляции.

Если обнаруживается пространственная взаимозависимость между регионами, то можно построить регрессионные модели с учетом пространственных эффектов, позволяющие определить, какие факторы влияют на исследуемый показатель, каковы масштабы влияния.

1.2. Типы данных

В основе пространственного анализа лежит использование данных. ***Данные*** – поддающееся многократной интерпретации представление информации в формализованном виде, пригодном для передачи, связи, или обработки[3] (определение по ISO/IEC 2382–1:1993).

Статистический анализ проводится на наблюдениях, которые образуют ***генеральную совокупность*** (если включены все возможные наблюдения) или ***выборочную совокупность*** (если включена часть наблюдений генеральной совокупности, отобранных по определенному признаку). В генеральную совокупность населения планеты Земля входят все 8 млрд. чел., но при проведении исследования нас может интересовать какая-то ее часть, например, в возрасте 18 лет и старше или проживающие на определенной территории. При переходе от выборочной совокупности к генеральной совокупности важно обращать внимание на следующие моменты:

1) выводы исследования, сделанные по выборочной совокупности, не всегда можно распространить на генеральную совокупность;

2) генеральная совокупность не всегда может быть доступна для исследования (в силу организационных, финансовых ограничений);

3) ***ошибка репрезентативности*** возникает в том случае, когда отбор наблюдений в выборочную совокупность проводится не случайном образом.

Единица совокупности – это предел дробления объекта исследования, при котором сохраняются все свойства изучаемого процесса[4]. При проведении статистического анализа исследователь работает со статистическими признаками, которые могут быть записаны в виде данных.

---

[3] Приказ Федерального агентства по техническому регулированию и метрологии от 22 сентября 2016 г. N 1189-ст «О введении в действие межгосударственного стандарта» (ГОСТ 33707-2016 (ISO/IEC 2382:2015)). URL: https://base.garant.ru/71572028/ (дата обращения: 11.11.2022).

[4] *Елисеева И.И., Юзбашев М.М.* Общая теория статистики. М. 2004. С. 21.

Данные могут быть классифицированы по различным признакам:

1) по характеру выражения: количественные и качественные. Наиболее типичной классификацией данных является выделение количественных (дискретных и непрерывных) и качественных (порядковых и номинальных) данных (см. рис. 1.2).

Количественные данные выражаются числами. ***Дискретные*** показатели могут принимать только целочисленные значения. ***Непрерывные*** показатели могут принимать дробные значения.

Качественные данные (описательные) выражаются словесно. Качественные признаки подразделяются на ***номинальные*** (по которым нельзя ранжировать данные) и ***порядковые*** (по которым можно проранжировать единицы совокупности).

2) по способу измерения: первичные и расчетные[5].

Первичные данные характеризуют единицу совокупности в целом и представляют собой ***абсолютные величины***, то есть показатели, выражающие уровень, объем, размер явлений и процессов, имеющие единицы измерения (например, численность населения региона, число родившихся). ***Расчетные*** показатели не измеряются непосредственно, а рассчитываются (например, производительность труда).

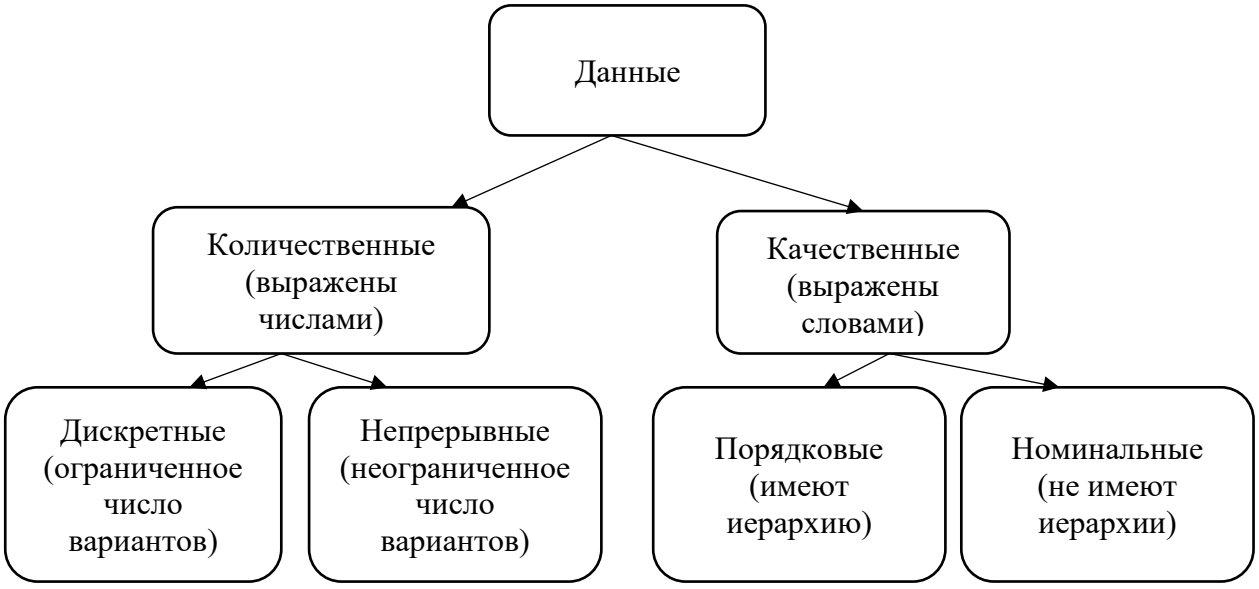

Рис. 1.2. Базовые типы данных

[5] Статистика / под редакцией И.И. Елисеевой. М., 2023. 361 с.

3) по структуре: перекрестные, временные и панельные данные (рис. 1.3). ***Перекрестные*** (cross-sectional) данные содержат значения исследуемого признака по *n* объектам. Примерами могут служить валовой региональный продукт по регионам России, численность населения по 45 муниципальным районам Республики Татарстан. Важной особенностью таких данных является привязка к конкретному периоду во времени, допустим, к году или определенной дате. ***Временные*** ряды содержат хронологически выстроенную последовательность значений исследуемого признака по одному объекту. Примером может быть динамика уровня инфляции в Республике Татарстан в 2010–2022 гг. В таком случае каждое значение признака привязано к определенному временному периоду и менять порядок данных нельзя, поскольку закономерность в динамике данного явления также может измениться. ***Панельные*** данные сочетают в себе черты перекрестных данных и временных рядов: они содержат значения показателей о развитии *n* объектов за *t* временных периодов.

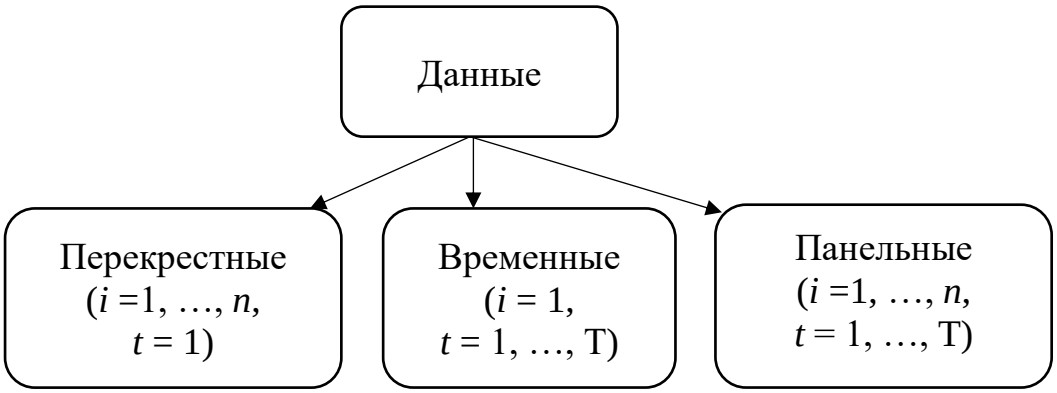

Рис. 1.3. Типы данных по структуре

Обращаем внимание на встречающуюся ситуацию, когда для перекрестных данных используют термин «пространственные». При этом в рамках настоящего пособия пространственные данные имеют более широкое значение. Они предполагают соотнесение собранных данных к показателям пространственной взаимосвязи.

Определенные правила применяются при вводе данных в статистические пакеты. Например, в GeoDa количественные данные могут быть записаны двумя способами:

1) целыми числами (integer);
2) действительными числами (real).

Качественные данные в GeoDa могут быть записаны тремя способами:

1) словами (string);
2) датами (date);
3) рангами.

Количественные и качественные данные дополнительно могут быть географически привязана к определенной территории.

Географически структурированные данные – это данные, которые помимо основной информации дополнительно содержат географические координаты (широта и долгота) или адреса объектов.

Географически структурированные данные являются одним из видов пространственно организованных данных.

Пространственно организованные данные – это данные, которые дополнительно содержат координаты, отражающие их привязку в пространстве.

Важно понимать, что в данном контексте понятие «пространство» используется в широком смысле и не обязательно сводится к географическому пространству. Пространство может быть экономическим, социальным, информационным и других видов. Например, в международной экономике расстояния между странами определяются товарооборотом, потоками капитала или мигрантов. В социологии расстояния между индивидуумами могут оцениваться длительностью или частотой общения, количеством друзей или подписчиков в социальных сетях.

Глоссарий

Абсолютные величины – показатели, выражающие уровень, объем, размер явлений и процессов, полученные в результате статистического наблюдения и имеющие единицы измерения.

Временные ряды содержат хронологически выстроенную последовательность значений исследуемого признака по одному объекту.

Выборочную совокупность – часть наблюдений генеральной совокупности, отобранных по определенному признаку.

Генеральная совокупность – совокупность объектов, в которую включены все возможные наблюдения.

Географически структурированные данные – это данные, которые помимо основной информации дополнительно содержат географические координаты (широта и долгота) или адреса объектов.

Географический детерминизм исходит из того, что расположение в пространстве является ведущим (иногда даже единственным) фактором, объясняющим свойства объекта.

Географический нигилизм предполагает, что географический фактор не влияет на свойства объекта.

Географический поссибилизм исходит из того, что пространство задает вероятностные сценарии для формирования свойства объекта, но не может однозначно определить его природу.

Данные – поддающееся многократной интерпретации представление информации в формализованном виде, пригодном для передачи, связи, или обработки.

Дискретные показатели могут принимать только целочисленные значения.

Единица совокупности – это предел дробления объекта исследования, при котором сохраняются все свойства изучаемого процесса

Качественные данные (описательные) выражаются словесно.

Количественные данные выражаются числами.

Непрерывные показатели могут принимать дробные значения.

Номинальные признаки – признаки, по которым нельзя ранжировать значения.

Ошибка репрезентативности возникает в том случае, когда отбор наблюдений в выборочную совокупность проводится не случайном образом.

Панельные данные сочетают в себе черты перекрестных данных и временных рядов: они содержат значения показателей о развитии *n* объектов за *t* временных периодов.

Первичные данные характеризуют единицу совокупности в целом.

Перекрестные (cross-sectional) данные содержат значения исследуемого признака по *n* объектам.

Порядковые признаки – статистические признаки, по которым можно проранжировать единицы совокупности.

Пространственно организованные данные – это данные, которые дополнительно содержат координаты, отражающие их привязку в пространстве.

Пространственный анализ – направление географических исследований, направленное на анализ свойств объекта, продиктованных его взаимосвязанностью с другими объектами.

Расчетные показатели не измеряются непосредственно, а рассчитываются.

Вопросы для самоконтроля

1. Дайте определение понятию «пространственный анализ». На какие исследовательские вопросы можно ответить, используя инструменты пространственного анализа?
2. Из каких этапов состоит пространственный анализ?
3. Какие ограничения необходимо учитывать при работе с выборочной совокупностью?
4. В чем отличие между качественными и количественными данными?
5. Приведите примеры номинальных и порядковых данных.

6. Приведите примеры дискретных и непрерывных показателей.

7. Как взаимосвязаны первичные и расчетные показатели?

8. Приведите примеры перекрестных, временных и панельных данных.

9. Какие данные называются географически структурированными? Приведите примеры.

10. Какие данные называются пространственно организованными? Приведите примеры из области географии, физики, химии, экономики, социологии, психологии.

ТЕМА 2. ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ ВЕРОЯТНОСТЕЙ И СТАТИСТИКИ, ПРИМЕНЯЕМЫЕ В ПРОСТРАНСТВЕННОМ АНАЛИЗЕ

В результате освоения темы обучающийся будет:

– ***знать*** основные понятия теории вероятности, виды распределений случайных величин;

– ***уметь*** определять вид распределения случайной величины;

– ***владеть*** навыками проверки статистических гипотез.

Основные вопросы:

2.1. Основные понятия теории вероятности.

2.2. Законы распределения случайной величины.

2.3. Проверка статистических гипотез.

Ключевые слова: случайная величина, вероятность случайной величины, закон распределения, нормальный закон распределения, распределение Стьюдента, χ^2-распределение, распределение Фишера, статистическая гипотеза, доверительный интервал, уровень значимости

2.1. Основные понятия теории вероятности

Теория вероятностей – отрасль математики, занимающаяся анализом случайных явлений. Исход случайного события не может быть определен до его наступления, но он может быть любым из нескольких возможных исходов. Фундаментальным элементом теории вероятности является эксперимент, который может быть повторен, по крайней мере гипотетически, при практически одинаковых условиях и который может привести к различным результатам при разных испытаниях. Набор всех возможных исходов эксперимента называется пространством выборки.

Событие – это четко определенное подмножество пространства выборки. Например, событие «сумма граней, выпавших на двух

игральных костях, равна шести» состоит из пяти исходов (1, 5), (2, 4), (3, 3), (4, 2) и (5, 1). Для идеализированного волчка, сделанного из отрезка прямой линии, не имеющего ширины и вращающегося в центре, множество возможных исходов – это множество всех углов, которые конечное положение волчка составляет с некоторым фиксированным направлением, эквивалентно всем действительным числам в [0, 2π). Многие измерения в естественных и общественных науках, такие как объем, напряжение, температура, время реакции, маржинальный доход и так далее, производятся на непрерывных шкалах и, по крайней мере теоретически, включают бесконечно много возможных значений. Если повторные измерения на разных субъектах или в разное время на одном и том же субъекте могут привести к различным результатам, теория вероятности является возможным инструментом для изучения этой изменчивости.

Согласно классической концепции, ***случайная величина*** (СВ) – это функция, которая отображает каждый результат эксперимента в пространстве выборок на числовое значение на вещественной прямой. Таким образом, СВ – это не переменная, а скорее функция. Случайная величина обычно обозначается заглавной буквой, например X, а соответствующая строчная буква, то есть x, используется для обозначения конкретного значения этой случайной величины. В зависимости от свойств различают дискретные и непрерывные случайные величины. Таким образом, если набор значений, которые может принимать случайная величина, конечен или счетно бесконечен, то говорят, что случайная величина является ***дискретной*** случайной величиной. Если множество значений, которые может принимать случайная величина, является континуумом, то есть все возможные значения в диапазоне действительных чисел, то случайная величина называется ***непрерывной*** случайной величиной.

Событиям присваивается вероятность, и это присвоение вероятности событиям является ключом для любой оценки. Для этого требуется концепция теории множеств, которая включает взаимосвязи

между событиями, такие как объединение (обозначается как $A \cup B$), пересечение (обозначается как $A \cap B$ или AB) и дополнение (обозначается как A^C). Графическое представление пространства выборки, событий и их взаимосвязей обычно изображается диаграммой Венна.

Два события E_1 и E_2 называются ***взаимоисключающими***, если ни один из исходов в E_1 не принадлежит E_2 или наоборот. Это обозначается как: $E_1 \cap E_2 = \emptyset$, где $\emptyset$ обозначает нулевое множество.

В любом случайном эксперименте всегда существует неопределенность в отношении того, произойдет ли конкретное событие или нет. Концепция вероятности была первоначально предложена для объяснения неопределенности, связанной с результатом случайного эксперимента.

В качестве измерения шанса или вероятности, с которой можно ожидать наступления события, удобно использовать число от 0 до 1. Согласно классическому определению, вероятность события A, обозначаемая как $P(A)$, определяется априори, без проведения реального эксперимента. ***Вероятность*** случайного события определяется соотношением:

$$P(A) = \frac{N_A}{N}, \tag{2.1}$$

где N – число возможных исходов, а N_A – число исходов, благоприятных для события A.

Определение вероятности в гидрологии и гидроклиматологии более эффективно выражается в терминах относительных частот[6]. Если случайное событие происходит большое число раз N и событие A происходит в n из этих случаев, то вероятность наступления события A равна:

$$P(A) = \lim_{N \to \infty} \frac{n}{N}. \tag{2.2}$$

[6] *Maity R.* Basic concepts of probability and statistics // Statistical Methods in Hydrology and Hydroclimatology. 2018. P. 7–51.

Вероятность любого события A в пространстве выборок S, обозначаемая как $P(A)$, назначается таким образом, чтобы она удовлетворяла определенным условиям. Эти условия для назначения вероятности известны как ***аксиомы вероятности***[7].

Аксиома 1. Вероятности для каждого события A в пространстве выборке S являются действительными числами в интервале от 0 до 1, включая границу, то есть 0 и 1:

$$0 \leq P(A) \leq 1. \tag{2.3}$$

Аксиома 2. Пространство выборки S в целом имеет вероятность, равную единице:

$$P(S) = 1. \tag{2.4}$$

Поскольку пространство выборки S содержит все возможные исходы, один из них всегда должен произойти.

Аксиома 3. Если A и B – взаимоисключающие события в пространстве выборки S, то вероятность этих событий будет равна сумме вероятностей:

$$P(A \cup B) = P(A) + P(B). \tag{2.5}$$

Это означает, что функции вероятности должны быть аддитивными, то есть вероятность объединения равна сумме двух вероятностей, когда два события не имеют общего исхода. Все выводы, сделанные по теории вероятностей, прямо или косвенно связаны с этими тремя аксиомами.

Напомним, что генеральной совокупностью называется вся совокупность объектов (наблюдений), которые являются объектом исследования. В математической статистике понятие генеральной совокупности трактуется как совокупность всех мыслимых наблюдений, которые могли бы быть произведены при данном реальном комплексе условий, и в этом смысле его не следует

[7] *Орлов А.И.* Вероятность и прикладная статистика – основные факты. М. 2010. 95 с.

смешивать с реальными совокупностями, подлежащими статистическому изучению[8].

Для проверки исследовательских гипотез не всегда бывает возможность работать со всеми наблюдениями генеральной совокупности, что может быть связано с временными, материальными, пространственными ограничениями по ресурсам. Следовательно, исследователи работают с частью наблюдений генеральной совокупности, отобранных определенным образом, – выборочной совокупностью.

2.2. Законы распределения случайной величины

Закон распределения случайной величины может быть представлен как соотношение, определяющее связь между значениями случайной величины и вероятностью их возникновения. Закон распределения можно представить таблично (см. табл. 2.1), в виде формулы или графически.

Таблица 2.1

Пример закона распределения случайной величины

X	10	20	30	100	200
p_i	0,25	0,25	0,20	0,15	0,15

Функцией распределения СВ X называют функцию $F(x)$, определяющую вероятность того, что СВ X принимает значение меньшее, чем x:

$$F(x) = P(X < x). \qquad (2.6)$$

[8] *Кремер Н.Ш.* Теория вероятностей и математическая статистика: учебник и практикум для академического бакалавриата. М., 2016. 514 с.

Плотностью вероятности (плотностью распределения вероятностей) непрерывной СВ *X* называется производная ее функции распределения[9]:

$$f(x) = F`(x). \quad (2.7)$$

Плотность вероятности *f(x)*, как и функция распределения *F(x)*, существует только для непрерывных случайных величин.

В рамках числовых характеристик выборочной совокупности можно выделить характеристики:

1) положения: математическое ожидание, мода, медиана, начальные моменты различных порядков;

2) рассеивания: дисперсия, среднее квадратическое отклонение, центральные моменты различных порядков[10].

Математическое ожидание представляет собой вероятностное среднее значение случайной величины. Для дискретной СВ:

$$M(x) = \sum_{i=1}^{n} x_i * p_{i.} \quad (2.8)$$

Для непрерывной величины:

$$M(x) = \int_{-\infty}^{+\infty} xf(x)dx. \quad (2.9)$$

Одной из наиболее популярных числовых характеристик выборочных совокупностей является ***выборочное среднее арифметическое***, которое представляет собой сумму значений наблюдений совокупности, деленную на объем выборки:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad (2.10)$$

[9] *Кремер Н.Ш., Путко Б.А.* Эконометрика: учебник для студентов вузов. М. 2012. 328 с.

[10] *Орлов А.И.* Вероятность и прикладная статистика – основные факты. М. 2010. 95 с.

где n – объем выборочной совокупности, x_i – значение i-ого наблюдения в выборке.

Дисперсией случайной величины X $D(X)$ называется математическое ожидание квадрата отклонения значений случайной величины от их математического ожидания:

$$D(X) = M\big(X - M(X)\big)^2. \qquad (2.11)$$

Среднее квадратическое отклонение случайной величины X рассчитывается как квадратный корень из дисперсии:

$$\sigma(x) = \sqrt{D(x)}. \qquad (2.12)$$

Относительным показателем, характеризующим неоднородность или однородность разброса значений выборочной совокупности относительно среднего значения, служит ***коэффициент вариации*** $V(x)$, который рассчитывается по формуле:

$$V(x) = \frac{\sigma(x)}{|M(x)|} * 100\ \%. \qquad (2.13)$$

Для прогнозирования будущих значений случайной величины, необходимо знать ее закон распределения, с помощью которого можно определить какое значение примет случайная величина с заданной вероятностью. В статистическом анализе наиболее часто используются следующие законы распределения:

1) нормальное распределение;
2) хи квадрат–распределение;
3) распределение Стьюдента;
4) распределение Фишера.

Если случайная величина имеет нормальный закон распределения, то в данной случае можно многое сказать о ее будущем поведении. Его

главная особенность, состоит в том, что он является предельным законом, к которому асимптотически приближаются другие законы распределения.

Непрерывная случайная величина X имеет ***нормальный закон распределения*** с математическим ожиданием m и дисперсией σ^2, если ее плотность вероятности может быть представлена в виде:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-m)^2}{2\sigma^2}}. \qquad (2.14)$$

Другими словами, говорят, что случайная величина является нормально распределенной или нормальной. График плотности вероятности нормальной случайной величины изображен на рис. 2.1.

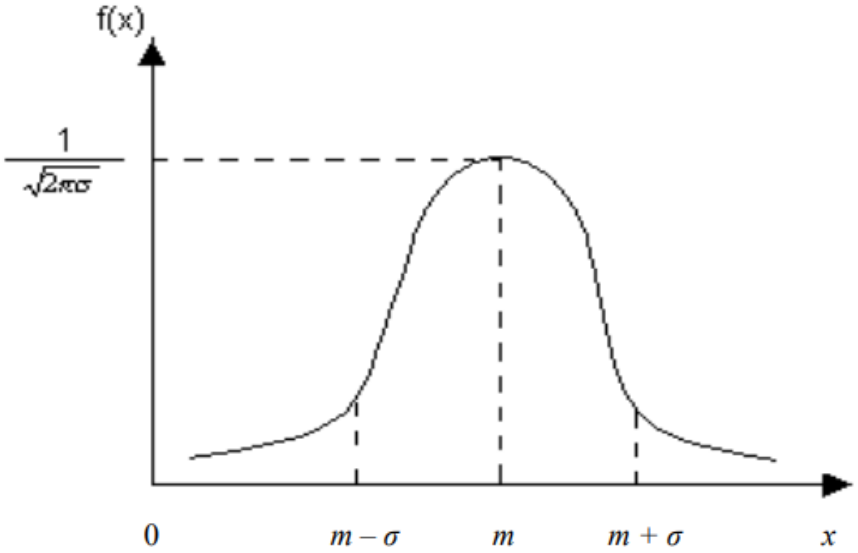

Рис. 2.1. График плотности вероятности нормального распределения случайной величины X

Частным случаем нормального распределения является ситуация, когда $m = 0$ и $\sigma = 1$. В этом случае говорят о ***стандартном нормальном распределении***.

Распределением χ^2 (хи-квадрат) с k степенями свободы называется распределение суммы квадратов k независимых случайных величин, распределенных по стандартному нормальному закону[11]:

[11] *Кремер Н.Ш., Путко Б.А.* Эконометрика: учебник для студентов вузов. М. 2012. 328 с.

$$\chi^2 = \sum_{i=1}^{k} U_i^2, \qquad (2.15)$$

где $U_i = \frac{(x_i - m_i)}{\sigma_i}$.

На рис. 2.2 показаны кривые χ^2-распределения для различных значений числа степеней свободы. Они показывают, что χ^2-распределение асимметрично, обладает положительной (правосторонней) асимметрией. При $k > 30$ распределение случайной величины $U = \sqrt{2\chi^2} - \sqrt{2k-1}$ близко к стандартному нормальному закону, то есть *N(0;1)*.

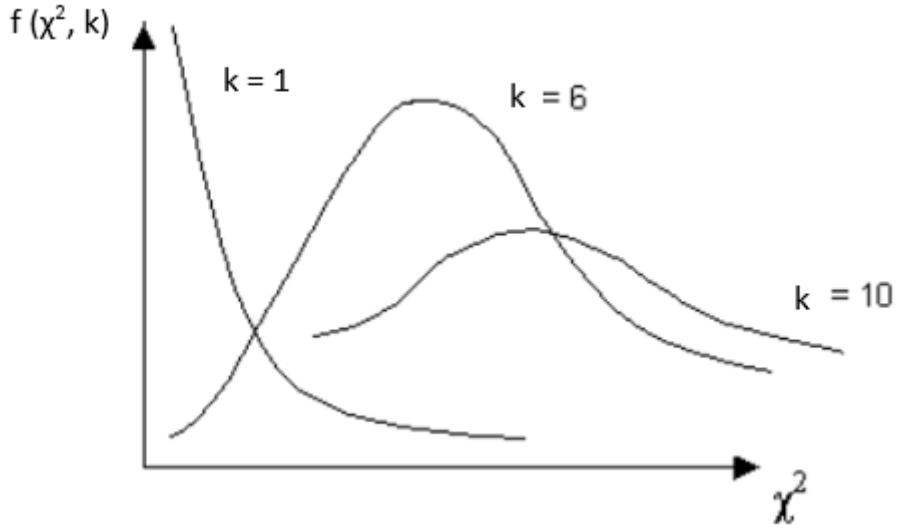

Рис. 2.2. Кривые χ^2-распределения для разных степеней свободы *k*

Распределением Стьюдента[12] (или *t*-распределением) называется распределение случайной величины, которое может быть представлено в виде:

$$t = \frac{Z}{\sqrt{\frac{1}{n} \cdot \chi^2}}, \qquad (2.16)$$

[12] *Кремер Н.Ш., Путко Б.А.* Эконометрика: учебник для студентов вузов. М. 2012. 328 с.

где Z – случайная величина, имеющая стандартный нормальный закон распределения, то есть $Z \sim N(0;1)$; χ^2 – независимая от Z случайная величина, имеющая χ^2-распределение с n степенями свободы. График плотности вероятности случайной величины X, имеющей распределение Стьюдента, при высоком числе степеней свободы n очень близко к нормальному распределению (см. рис. 2.3).

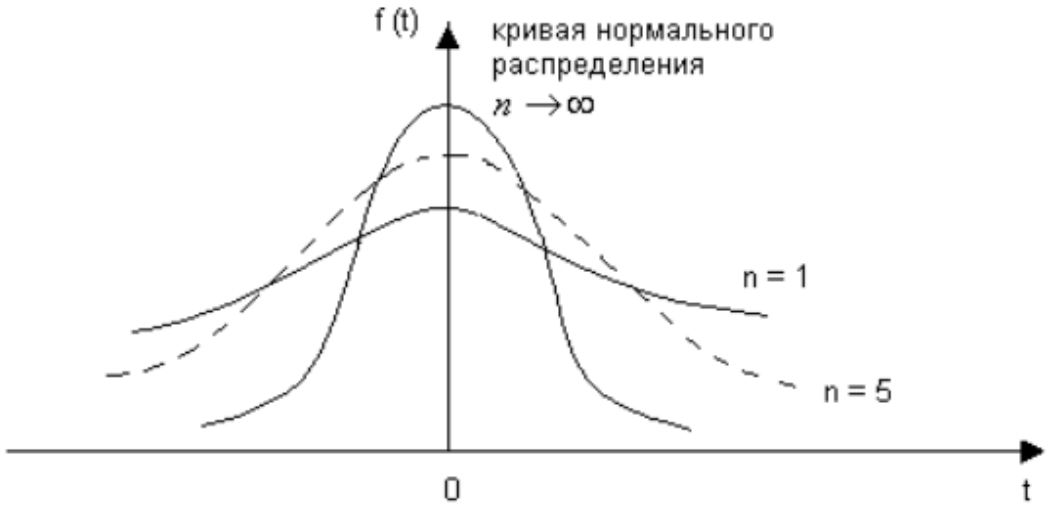

Рис. 2.3. График функции плотности вероятности случайной величины X, имеющей распределение Стьюдента с n степенями свободы

Распределением Фишера-Снедекора (или F-распределением) называется распределение случайной величины, которая может быть представлена в виде:

$$F = \frac{\frac{1}{m} \cdot \chi^2(\mathrm{m})}{\frac{1}{n} \cdot \chi^2(\mathrm{n})}, \qquad (2.17)$$

где $\chi^2(m)$ и $\chi^2(n)$ – случайные величины, имеющие χ^2-распределение соответственно с m и n степенями свободы.

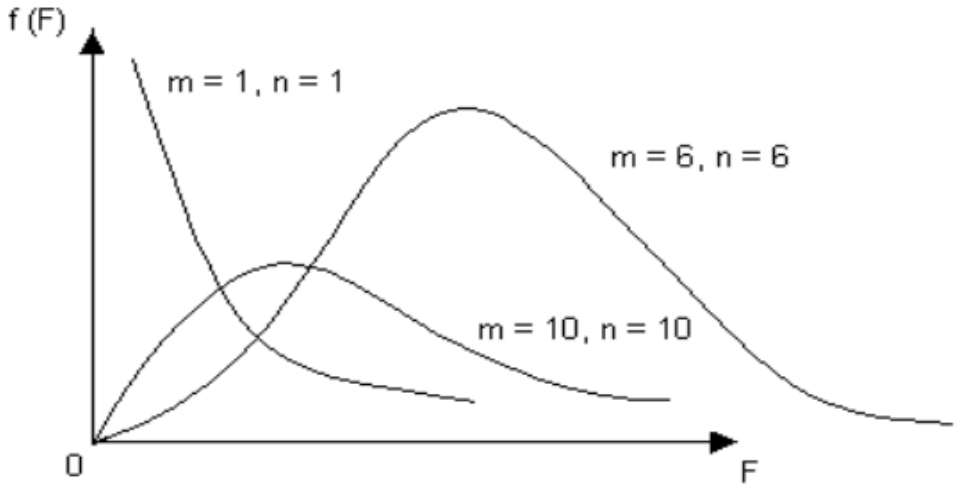

Рис. 2.4. Кривые *F*-распределения
при некоторых значениях числа степеней свободы *m* и *n*

Нормальное распределение, χ^2-распределение, распределение Стьюдента и распределение Фишера используются при проверке статистических гипотез в дисперсионном и регрессионном анализах. Для них разработаны специальные таблицы критических точек распределения, значения которых необходимы для корректного вывода о принятии или отклонении статистических гипотез.

2.3. Проверка статистических гипотез

Статистический анализ предполагает приближенное определение неизвестных распределений вероятностей или их отдельных числовых характеристик (например, математического ожидания).

Результатом ***точечного оценивания*** является приблизительное значение случайной величины, которое может быть представлено в виде одного числа. К основным методам точечного оценивания относятся: метод максимального правдоподобия, метод моментов, метод наименьших квадратов. Получаемые в ходе анализа точечные оценки априори не являются точными, поэтому они должны характеризоваться такими свойствами, как: несмещенность, эффективность, состоятельность.

Несмещенность оценки предполагает отсутствие систематического смещения значения показателя относительно его истинного значения в генеральной совокупности. Эффективность точечной оценки продиктовано возможностью, когда, используя различные статистические или эконометрические методы, можно найти несколько несмещенных оценок для одного и того же параметра. В таком случае возникает вопрос – какая из полученных оценок лучше? Критерием сравнения полученных оценок является наименьший разброс значений, полученных с помощью оценки, иными словами – наименьшая дисперсия, то есть чем меньше дисперсия полученных значений, тем выше эффективность оценки. Следовательно, ***эффективной*** называется несмещенная оценка с минимальной дисперсией. ***Состоятельной*** называется оценка, значение которой с увеличением объема выборки приближается к истинному значению параметра генеральной совокупности.

Точечное оценивание не всегда представляется информативным с точки зрения того, что получение одного единственного значения показателя в условиях неизвестной информации выглядит неубедительным. Соответственно, ***интервальное оценивание*** позволяет оценить неизвестное значение параметра в виде интервала его допустимых значений и определения вероятности того, что в этом интервале находится истинное значение параметра. ***Доверительным интервалом*** для параметра генеральной совокупности называется отрезок, содержащий истинное значение данного параметра с большой вероятностью. Границы доверительного интервала оцениваются по выборочной совокупности. Если много раз из генеральной совокупности брать независимые выборки и по каждой из них строить доверительный интервал для исследуемого параметра, то определенная доля этих интервалов будет содержать внутри себя значение параметра. Доверительный интервал строят так, чтобы доля накрывающих интервалов равнялась доверительному уровню (уровню доверительной надежности). Стандартные значения доверительных уровней: 95 %, 90 %, 99 % и, реже, 99,9 %. Ширина

доверительного интервала характеризует степень неопределенности: чем шире доверительный интервал, тем меньше точной информации мы можем сказать о значении данного параметра. Тем не менее, доверительные интервалы дают больше информации о параметре, чем простая точечная оценка, поскольку ограничивают совокупность допустимых значений.

Статистической гипотезой называется выдвигаемое предположение о виде или параметрах неизвестного закона распределения. Проверяемую гипотезу обычно называют нулевой (или основной) и обозначают H_0. Наряду с нулевой гипотезой H_0 рассматривают альтернативную, или конкурирующую, гипотезу H_1 являющуюся логическим отрицанием H_0. Нулевая и альтернативная гипотезы представляют собой две возможности выбора, осуществляемого в задачах проверки статистических гипотез[13].

Проверка статистической гипотезы предполагает последовательное выполнение следующих этапов[14]:

1. Описание статистической модели выборочной совокупности.

2. Формулировка нулевой и альтернативной гипотез.

3. Установление уровня значимости, с помощью которого контролируется ошибка 1-го рода. ***Уровень статистической значимости (α)*** – это вероятность ошибочного отклонения нулевой гипотезы. ***Ошибка 1-го рода*** – это ситуация, когда отвергнута верная нулевая гипотеза. Уровень значимости тесно связан с доверительной вероятностью (γ):

$$\gamma = 1 - \alpha. \qquad (2.18)$$

4. Выбор критерия для проверки нулевой гипотезы. Применение критерия позволит контролировать вероятность появления ошибки 2-

[13] *Кремер Н.Ш., Путко Б.А.* Эконометрика: учебник для студентов вузов. М. 2012. 328 с.

[14] *Кремер Н.Ш.* Теория вероятностей и математическая статистика: учебник и практикум для вузов. М. 2012. 551 с.

го рода. ***Ошибка 2-го рода*** – это ситуация, когда принята неверная нулевая гипотеза.

5. Вычисление фактического значения критерия в соответствии с определенным алгоритмом проверки данного рода гипотез исходя из имеющейся исходной информации.

6. Определение критической области и области согласия с нулевой гипотезой, то есть установление табличного значения критерия.

7. Сопоставление фактического и табличного значений критерия и формулирование выводов по результатам проверки нулевой гипотезы.

Глоссарий

Вероятность случайного события определяется соотношением благоприятных исходов к общему количеству исходов.

Взаимоисключающими называются два события E_1 и E_2, если ни один из исходов в E_1 не принадлежит E_2 или наоборот.

Выборочное среднее арифметическое – сумма значений случайной величины, деленная на количество наблюдений.

Дисперсия случайной величины X $D(X)$ – это математическое ожидание квадрата отклонения случайной величины от ее математического ожидания.

Доверительный интервал для параметра генеральной совокупности – это отрезок, с большой вероятностью содержащий этот параметр.

Закон распределения случайной величины может быть представлен как соотношение, определяющее связь между значениями случайной величины и вероятностью их возникновения.

Интервальное оценивание – способ получения оценки для неизвестного значения параметра с помощью интервала его допустимых значений и определения вероятности того, что в этом интервале находится истинное значение параметра.

Математическое ожидание характеризует среднее ожидаемое значение случайной величины.

Несмещенность оценки предполагает отсутствие систематического смещения значения показателя относительно его истинного значения в генеральной совокупности.

Нормально распределенная (или нормальная) случайная величина – это случайная величина, имеющая нормальное распределение.

Ошибка 1-го рода – это ситуация, когда отвергнута верная нулевая гипотеза.

Ошибка 2-го рода – это ситуация, когда принята неверная нулевая гипотеза.

Плотность вероятности (плотностью распределения вероятностей) непрерывной случайной величины X – это производная ее функции распределения.

Распределением χ^2 (хи-квадрат) с k степенями свободы называется распределение суммы квадратов k независимых случайных величин, распределенных по стандартному нормальному закону.

Случайная величина – это функция, которая отображает каждый результат эксперимента в пространстве выборок на числовое значение на вещественной прямой.

Событие – это четко определенное подмножество пространства выборки.

Состоятельной называется оценка, значение которой с увеличением объема выборки приближается к истинному значению параметра генеральной совокупности.

Среднее квадратическое отклонение случайной величины X – это квадратный корень из дисперсии.

Стандартное нормальное распределение – это частный случай нормального распределения с нулевым математическим ожиданием и дисперсией, равной 1.

Статистической гипотезой называется выдвигаемое предположение о виде или параметрах неизвестного закона распределения.

Точечное оценивание предполагает получение приблизительного значения параметра в виде одного числа.

Уровень статистической значимости (α) – это вероятность ошибочного отклонения верной нулевой гипотезы.

Функцией распределения случайной величины X называют функцию $F(x)$, определяющую вероятность того, что случайная величина X принимает значение меньшее, чем x.

Эффективной называется несмещенная оценка с минимальной дисперсией.

Вопросы для самоконтроля

1. Зачем необходимы знания по теории вероятности и математической статистике для проведения статистического анализа пространственных данных?
2. Что называется случайной величиной? Приведите примеры случайных величин в пространственном анализе.
3. Что такое вероятность случайной величины? Как ее найти?
4. Что такое закон распределения случайной величины? Для чего его нужно знать?
5. Какие законы распределения используются для проверки статистических гипотез?
6. Какими свойствами должна обладать точечная оценка?

ТЕМА 3. ОПИСАТЕЛЬНЫЕ СТАТИСТИКИ ДЛЯ АНАЛИЗА ДВУХ И БОЛЕЕ ПЕРЕМЕННЫХ. КОРРЕЛЯЦИЯ

В результате освоения темы обучающийся будет:

– ***знать*** основные показатели описательной статистики для анализа двух и более переменных;

– ***уметь*** диагностировать наличие или отсутствие корреляции между переменными;

– ***владеть*** навыками корреляционного анализа.

Основные вопросы:

3.1. Корреляция: понятие, интерпретация.

3.2. Коэффициенты корреляции.

3.3. Корреляционный анализ.

Ключевые слова: корреляция, казуальность, шкала Чеддока, корреляционный анализ, коэффициент линейной корреляции Пирсона, коэффициент корреляции Кендалла, коэффициент корреляции Спирмена

3.1. Корреляция: понятие, интерпретация

Явления, процессы общественной жизни существуют не изолированно, а находятся в постоянном взаимодействии и взаимовлиянии с другими событиями. Когда мы переходим от анализа одной переменной, к исследованию взаимосвязи с другими переменными, нам необходимо проверить, существует ли подобная взаимосвязь и оценить ее силу.

Для анализа двух или более переменных используется двумерная описательная статистика – корреляция. ***Корреляция*** – это взаимная связь между двумя или более двух случайными величинами. Сущность корреляции заключается в том, что при изменении значе-

ния одной переменной относительно среднего значения происходит изменение (уменьшение или увеличение) другой переменной относительно ее среднего значения.

Корреляционный анализ является начальной стадией статистического анализа. Он позволяет понять, существует ли взаимосвязь между исследуемыми показателями или она отсутствует. Например, исследовательскими вопросами, на которые можно ответить с помощью оценки корреляции, являются следующие: существует ли взаимосвязь между безработицей и уровнем инфляции, между ростом студента и его академическим рейтингом, между количеством лет образования индивида и уровнем дохода.

Важно понимать, что корреляция отражает только взаимосвязь между переменными и не говорит о причинно-следственных связях, то есть ***каузальности***. Например, если по выборке индивидов обнаружена корреляция между ростом и весом человека, то это бы означало, что при увеличении веса человека его рост должен был бы увеличиваться (уменьшаться).

Корреляционные связи традиционно классифицируются по следующим критериям: форма, направление и степень (сила). По форме корреляционная связь может быть линейной или нелинейной. Примером линейной связи служит связь между количеством выполненных домашних заданий по теме и количеством правильно решаемых задач в контрольной работе. Нелинейной может быть взаимосвязь между уровнем стресса и производительностью труда. При повышении уровня стресса производительность труда сначала возрастает, затем достигается оптимальный уровень производительности труда; дальнейшему повышению уровня стресса будет соответствовать уже снижение производительности труда (см. рис. 3.1).

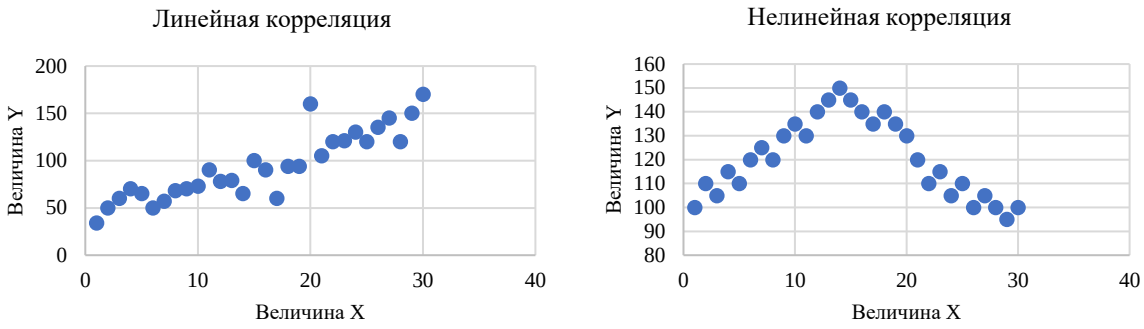

Рис. 3.1. Линейная и нелинейная корреляция

Когда корреляция ***прямая положительная***, то существенным отклонениям значений одного показателя от среднего значения соответствуют такие же отклонения значений другого показателя от его среднего, то есть отклонения – однонаправлены. При положительной корреляции коэффициент имеет положительный знак (например, r_{xy} = 0,7). При ***отрицательной*** корреляции направления изменений – противоположные, и коэффициент имеет отрицательный знак (например, r_{xy} = -0,7) (см. рис. 3.2).

Силу корреляционной связи можно определить по величине коэффициента корреляции. Она не зависит от её направленности и определяется по абсолютному значению коэффициента корреляции. Абсолютное значение линейного коэффициента корреляции может варьироваться в интервале от нуля до единицы.

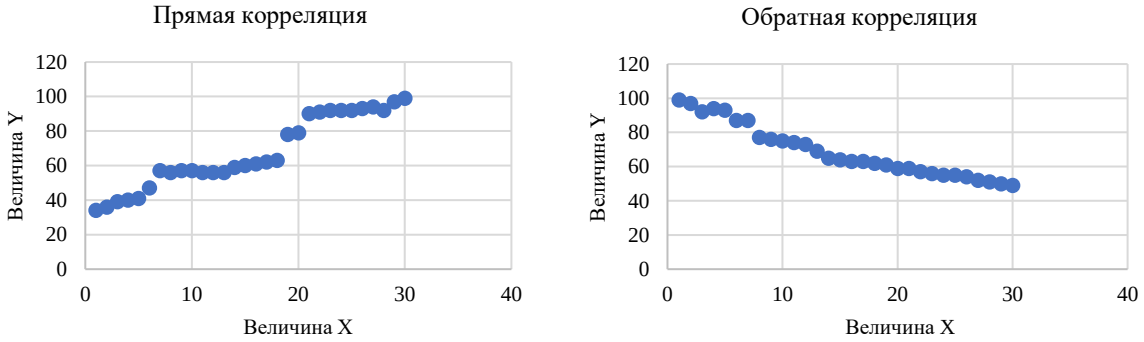

Рис. 3.2. Прямая и обратная корреляция

Корреляционную связь классифицируют на несколько групп (так называемая ***шкала Чеддока***): сильная или тесная при коэффициенте корреляции ($|r_{xy}| > 0{,}70$); средняя (при $0{,}50 < |r_{xy}| < 0{,}69$); умеренная (при $0{,}30 < |r_{xy}| < 0{,}49$); слабая (при $0{,}20 < |r_{xy}| < 0{,}29$); очень слабая (при $|r_{xy}| < 0{,}19$)[15]. Графически слабая корреляция представляет собой разреженное облако точек (см. рис. 3.3).

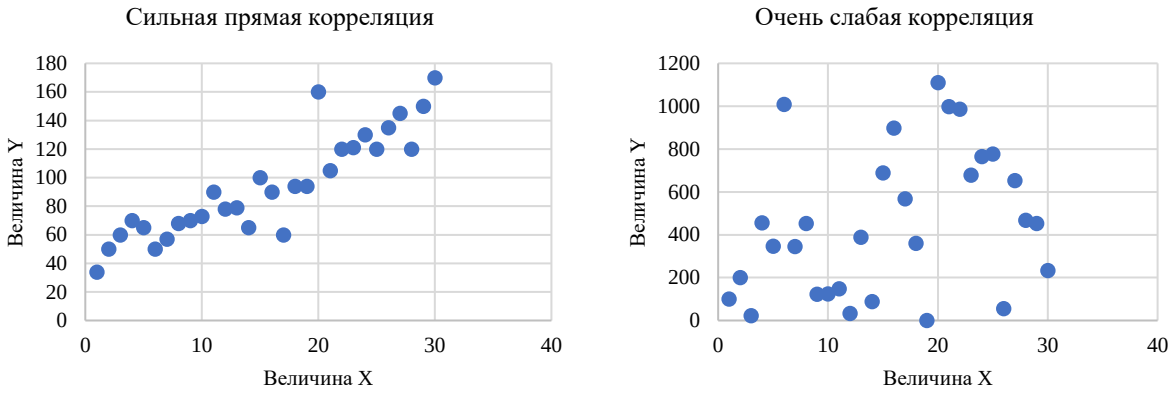

Рис. 3.3. Виды корреляции по степени

Визуальная интерпретация взаимосвязи между переменными не всегда бывает очевидной, поэтому для более четкого определения силы взаимосвязи между показателями рассчитывают специальные коэффициенты, имеющие более четкую интерпретацию.

3.2. Коэффициенты корреляции

На сегодняшний день разработано достаточно много различных коэффициентов корреляции. Остановимся более подробно на коэффициентах корреляции Пирсона, Кендалла и Спирмена. Все три коэффициента показывают взаимную связь двух признаков, которые измерены в количественной шкале – ранговой или метрической.

[15] *Орлов А.И.* Вероятностно-статистические модели корреляции и регрессии // Научный журнал КубГАУ. 2020. № 160. С. 1–3.

Линейный коэффициент корреляции Пирсона можно использовать для оценки линейной взаимосвязи между двумя показателями, характеризующими признаки наблюдений одной и той же выборки. Данный коэффициент показывает, насколько пропорциональна изменчивость двух переменных. Следует помнить, что коэффициент корреляции Пирсона работает только в отношении линейной взаимосвязи между переменными, для случаев нелинейной взаимосвязи он работать не будет и примет значение, близкое к нулю.

Предпосылками для расчета коэффициента корреляции Пирсона выступают следующие моменты:

– предполагается, что исследуемые переменные X и Y имеют нормальное распределение;

– переменные X и Y должны быть измерены в интервальной шкале или шкале отношений;

– количество значений у переменных X и Y должно быть одинаковым.

При расчете коэффициента линейной корреляции Пирсона используется следующая формула[16]:

$$r_{xy} = \frac{\sum(x_i - \overline{x}) * (y_i - \bar{y})}{\sqrt{\sum(x_i - \overline{x})^2} * \sqrt{\sum(y_i - \bar{y})^2}}, \qquad (3.1)$$

где x_i – значения переменной X; y_i – значения переменной Y; $\bar{x}$ – среднее арифметическое для переменной X; $\bar{y}$ – среднее арифметическое для переменной Y.

Абсолютное значение линейного коэффициента корреляции варьируется от нуля до одного.

Недостатками коэффициента являются следующие моменты:

1. Неустойчивость к выбросам.

2. С помощью этого коэффициента корреляции можно найти лишь линейную взаимосвязь между переменными.

---

[16] *Магнус Я.Р., Катышев П.К., Пересецкий А.А.* Эконометрика. Начальный курс. М. 2021. 504 с.

Коэффициент ранговой корреляции Спирмена применяется для исследования корреляционной взаимосвязи между двумя ранговыми переменными.

Коэффициент ранговой корреляции Спирмена можно найти двумя способами:

1. При помощи формулы коэффициента корреляции Пирсона в случае, когда переменные заранее проранжированы.

2. Использование упрощенной формулы коэффициента корреляции, когда нет повторяющихся рангов:

$$r_s = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)}. \qquad (3.2)$$

Коэффициент ранговой корреляции Кендалла используется в случае нелинейной взаимосвязи между двумя переменными. Метод расчета строится на присвоение рангов исходным значениям переменных. Далее определяется количество совпадений и количество инверсий в рангах. Коэффициент ранговой корреляции Кендалла строится как разность между количеством совпадений и инверсий в рангах. Как правило, для одной и той же выборки коэффициент корреляции Спирмена будет принимать значение, превышающее значение коэффициента ранговой корреляции Кендалла. При этом уровень значимости будет совпадать.

Формула вычисления коэффициента ранговой корреляции Кендалла:

$$r = \frac{P(p) - P(q)}{n\dfrac{(n-1)}{2}}, \qquad (3.3)$$

где *P(p)* – число совпадений, *P(q)* – число инверсий, *n* – объем выборки.

При наличии связанных рангов формула изменяется с учетом поправки на связанные ранги:

$$r = \frac{P(p) - P(q)}{\sqrt{(\left[n\frac{(n-1)}{2}\right] - K_x) * (\left[n\frac{(n-1)}{2}\right] - K_y)}}, \qquad (3.4)$$

где $P(p)$ – число совпадений, $P(q)$ – число инверсий, n – объем выборки, K_x – поправка на связи рангов переменной X, K_y – поправка на связи рангов переменной Y[17].

3.3. Корреляционный анализ

Корреляционный анализ обычно предшествует этапу построению регрессионной модели и преследует своей целью выявление статистически значимых взаимосвязей между исследуемыми показателями. Корреляционный анализ позволяет определить наличие, форму, направление и силу связи между зависимой и независимыми переменными, между независимыми переменными. На этапе корреляционного анализа можно выявить сильную взаимосвязь между независимыми переменными и избежать в дальнейшем проблему мультиколлинеарности при построении регрессионных моделей.

Проверка гипотезы о значимости коэффициента корреляции Пирсона проводится традиционным для статистических тестов способом. Допустим по выборочной совокупности, состоящей из n наблюдений, был рассчитан выборочный коэффициент корреляции r_B. Нужно при заданном уровне значимости α проверить нулевую гипотезу о равенстве нулю коэффициента корреляции в генеральной совокупности[18]:

$$H_0: r_\text{г} = 0. \qquad (3.5)$$

---

[17] *Жаворонков А.В., Королёв А.Л.* Результаты применения коэффициентов корреляции Кендалла для выявления определенных параметров // Модернизация отечественной системы управления: анализ тенденций и прогноз развития: материалы Всероссийской научно-практической конференции и XII-XIII Дридзевских чтений. М., 2014. С. 191–196.

[18] *Черненко В.Д.* Высшая математика в примерах и задачах: учебное пособие для вузов. СПб. 2011. 507 с.

Для проверки гипотезы рассчитывается случайная величина:

$$T = \frac{r_B\sqrt{n-2}}{\sqrt{1-r_B}}, \quad (3.6)$$

имеющая распределение Стьюдента с $k = n - 2$ степенями свободы. Критическая область является двусторонней и задается как: $|T| > t_{\text{кр.}}$, $t_{\text{кр.}}(\alpha/2, k)$.

Следует обратить внимание, что для правомерности проверки значимости коэффициента линейной корреляции Пирсона должно соблюдаться условие о том, что случайные величины *X* и *Y* имеют нормальное распределение, что редко встречается для реальных данных[19].

В ходе анализа можно визуально посмотреть на связь различных переменных одновременно. Для этого можно использовать различные встроенные инструменты корреляционного анализа:

1. Матрица графиков.

Матрица диаграмм рассеяния – таблица (или матрица) точечных диаграмм, которые используются для отображения двумерных отношений между комбинациями числовых переменных. Каждая точечная диаграмма матрицы отображает отношение между двумя переменными, что позволяет на одном графике показать много отношений[20].

Матрица диаграмм рассеяния состоит из таблицы мини-графиков и одного графика предпросмотра, отображающего выбранный мини-график в большей детальности. Кроме того, в матрицу можно добавить гистограмму с распределением каждой числовой переменной. На рисунке 3.4 представлена матрица по двум показателям: уровень безработицы и среднемесячная номинальная начисленная заработная плата работников организации, скорректированная на стои-

[19] *Орлов А.И.* Вероятностно-статистические модели корреляции и регрессии // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2020. № 160. С. 130–162.

[20] *ESRI.* Матрица диаграммы рассеяния. URL: https://clck.ru/32nSN4 (дата обращения: 27.11.2022).

мость фиксированного набора товаров и услуг, по регионам России за 2019 г.

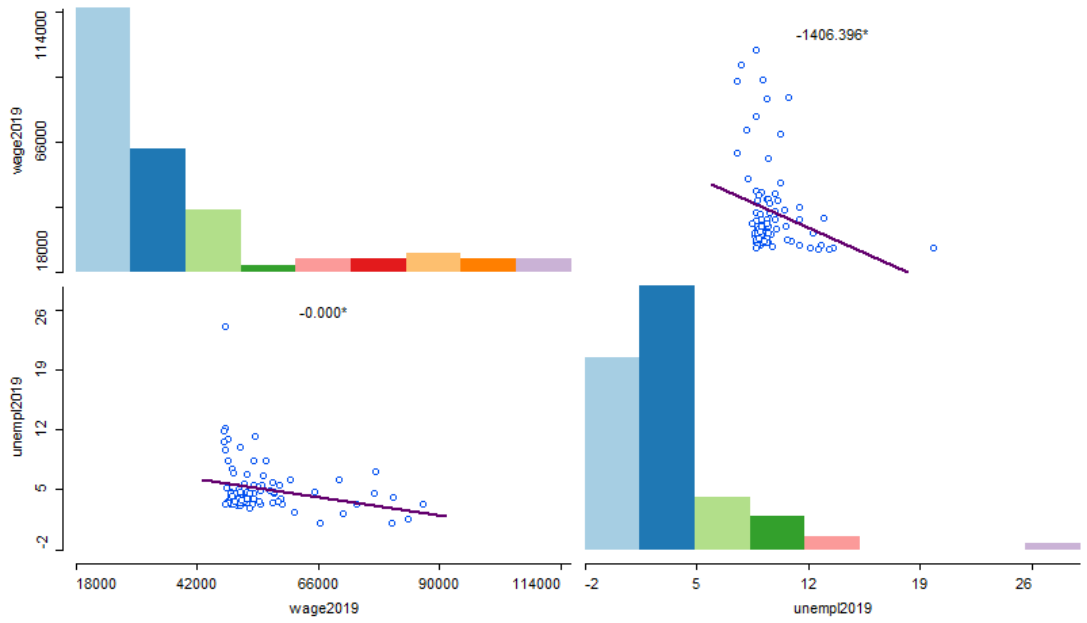

Рис. 3.4. Матрица диаграмм рассеяния

2. ***Матрица коэффициентов корреляции*** показывает одновременно несколько попарно вычисленных коэффициентов линейной корреляции (см. рис. 3.5).

```
Коэффициенты корреляции, наблюдения 1 - 85
5% критические значения (двухсторонние) = 0,2133 для n = 85

       Y          X1          X2          X3          X4
  1,0000     -0,2868     -0,2956      0,6766     -0,2597   Y
             1,0000      0,2256     -0,2215      0,1328   X1
                         1,0000     -0,3870      0,0773   X2
                                     1,0000      0,1283   X3
                                                 1,0000   X4
```

Рис. 3.5. Матрица коэффициентов корреляции

В качестве визуального инструмента также удобно использовать корреляционную матрицу, которая показывает силу корреляции между двумя переменными (см. рис. 3.6).

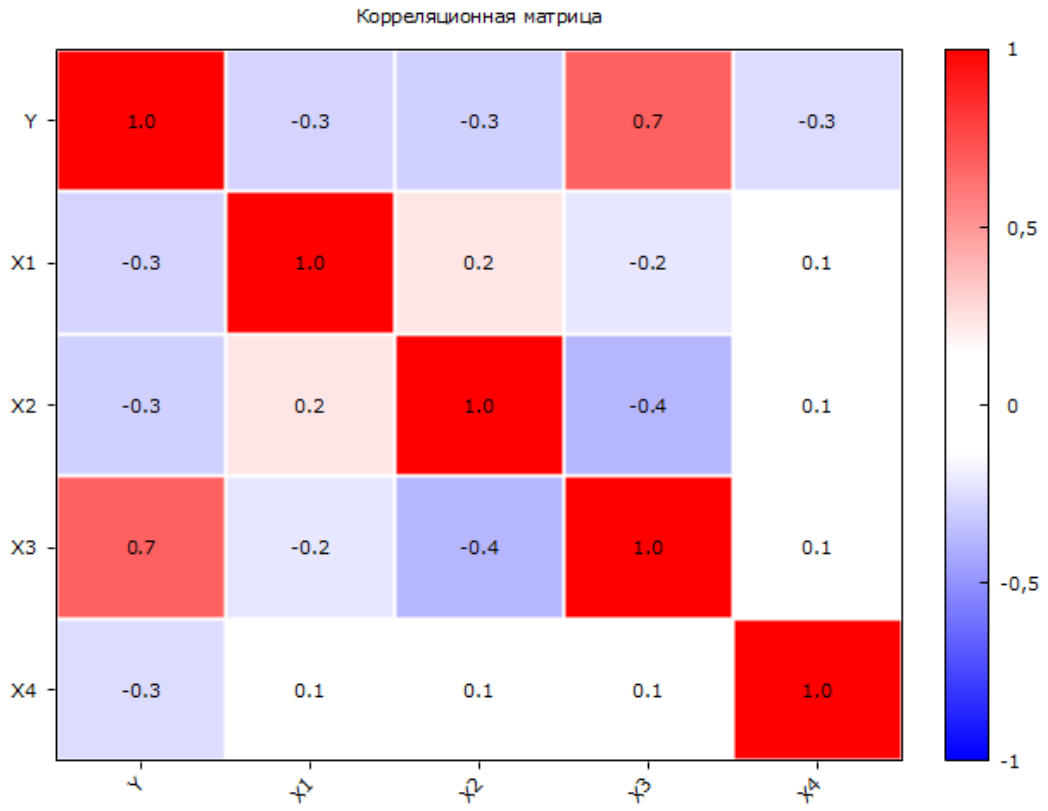

Рис. 3.6. Корреляционная матрица

Корреляционный анализ проводится на этапе проверки гипотезы о наличии взаимосвязи между исследуемыми переменными. При этом важно помнить о наличии ограничений при использовании тех или иных инструментов корреляционного анализа.

Глоссарий

Каузальность – причинно-следственная связь.

Корреляционный анализ позволяет определить наличие, форму, направление и силу связи между зависимой и независимыми переменными, между независимыми переменными.

Корреляция – это взаимная связь двух случайных величин.

Коэффициент ранговой корреляции Спирмена оценивает нелинейную взаимосвязь между двумя показателями.

Коэффициент ранговой корреляции Кендалла предназначен для определения взаимосвязи между двумя ранговыми переменными.

Линейный коэффициент корреляции Пирсона оценивает линейную взаимосвязь между двумя показателями.

Матрица коэффициентов корреляции показывает одновременно несколько попарно вычисленных коэффициентов линейной корреляции.

Отрицательная корреляция наблюдается в том случае, когда соотношения между изменениями двух переменных обратные и коэффициент корреляции имеет отрицательный знак.

Прямая положительная корреляция наблюдается в том случае, когда изменениям значений одного показателя соответствуют однонаправленные изменения значений другого показателя.

Шкала Чеддока позволяет классифицировать корреляционную связь по силе.

Вопросы для самоконтроля

1. В чем отличие корреляции от каузальности?
2. Какие бывают виды корреляции?
3. Какие коэффициенты корреляции относятся к параметрическим, а какие – к непараметрическим?
4. Какие условия должны выполняться, чтобы линейный коэффициент корреляции показывал достоверные результаты?
5. Если коэффициент выборочный коэффициент корреляции равен нулю, можно ли утверждать, что связь между переменными отсутствует?
6. Какие этапы включает в себя корреляционный анализ? Для каких целей он проводится?
7. С помощью каких инструментов можно определить наличие корреляции между несколькими переменными?

ТЕМА 4. ЛИНЕЙНАЯ РЕГРЕССИЯ В МОДЕЛИРОВАНИИ ПРОСТРАНСТВЕННЫХ ЗАВИСИМОСТЕЙ

В результате изучения данной темы обучающийся будет:

– ***знать*** определение теоретической и выборочной парной регрессии; предположения, лежащие в основе линейной регрессии; формулы для нахождения МНК-оценок линейной регрессии и его предпосылки; баланс сумм квадратов отклонений; определение и интерпретацию коэффициента детерминации;

– ***уметь*** находить оценки МНК-коэффициентов парной регрессии; определять качество подгонки линейной регрессии с помощью коэффициента детерминации, статистики Фишера и Стьюдента; интерпретировать графики остатков, выявлять влиятельные точки;

– *владеть* навыками оценки линейной регрессии; навыками интерпретации диаграмм рассеяния, статистик и диагностических метрик для оценки регрессионной модели.

Основные вопросы:

4.1. Спецификация линейной регрессии.

4.2. Метод наименьших квадратов и его предпосылки.

4.3. Коэффициент детерминации.

4.4. Проверка статистической значимости уравнения и коэффициентов регрессии.

4.5. Стандартизованные остатки регрессии.

Ключевые слова: линейная регрессия, метод наименьших квадратов, коэффициент регрессии, коэффициент детерминации, тест Фишера, тест Стьюдента, точка выброса

4.1. Спецификация линейной регрессии

Линейная регрессия – это метод статистического анализа, выявляющий линейные отношения между зависимой переменной y, и одной или несколькими независимыми переменными x. В зависимости от количества факторов, включенных в уравнение регрессии, принято различать простую (парную) и множественную регрессии. ***Парная регрессия*** представляет собой модель вида:

$$\hat{y} = f(x), \tag{4.1}$$

где $\hat{y}$ – предсказанное (теоретическое, расчетное) значение зависимой переменной (регрессанта, результативного признака); x – независимая, или объясняющая переменная (регрессор).

Любое эконометрическое исследование пространственных зависимостей начинается со спецификации модели, т. е. с формулировки вида модели, исходя из соответствующей теории связи между переменными. Из всего круга показателей, влияющих на регрессант, необходимо выделить влияющие наиболее существенно. Парная регрессия достаточна, если имеется доминирующий фактор, который и используется в качестве объясняющей переменной. Линейный регрессионный анализ основан на подборе прямой линии, наилучшим образом аппроксимирующей набор данных, с целью создания единственного уравнения, описывающего набор данных. Для предсказанных (теоретических) значений регрессора ***эмпирическое уравнение*** линейной парной регрессии имеет следующий вид:

$$\hat{y} = a + bx, \tag{4.2}$$

где a – выборочная оценка свободного коэффициента – точки пересечения с осью Y; b – выборочная оценка коэффициента регрессии, определяющего наклон линии регрессии.

Для фактических выборочных значений регрессора эмпирическая ***вероятностная модель*** линейной парной регрессии имеет вид:

$$y = a + bx + e, \tag{4.3}$$

где y – фактическое значение зависимой переменной, e – остаток (ошибка) регрессии, $e = y - \hat{y}$.

Гипотетически для значений регрессора в генеральной совокупности наблюдений можно записать ***теоретическую модель*** линейной парной регрессии:

$$y = \alpha + \beta x + \varepsilon, \tag{4.4}$$

где α – свободный коэффициент, β – коэффициент регрессии, ε – случайное отклонение (ошибка).

Спецификация теоретической модели линейной парной регрессии используется в постановке гипотез для верификации выборочных оценок.

Очевидно, что вероятностная модель состоит из детерминированного компонента $\hat{y}$ и компонента случайной ошибки e[21]. Каждое наблюдение зависимой переменной является суммой предсказанного значения и остатка: $y = \hat{y} + e$. Поэтому линия регрессии приближенно описывает облако точек, предсказанное значение зависимой переменной не совпадает с фактическим значением, которое в реальности наблюдается в данных. ***Остаток регрессии e*** измеряет влияние не учтенных в модели регрессоров, случайных ошибок и особенностей измерения. Его порождают 3 источника: спецификация модели, выборочный характер исходных данных и ошибки измерения (см. рис. 4.1).

Геометрически остатки регрессии – это отрезки по вертикали между реальными наблюдениями и предсказанными (линия регрессии). Положительные остатки – сверху линии регрессии, отрицательные – снизу. И сумма остатков равна нулю. Поэтому вместо суммы

[21] *Грекусис Дж.* Методы и практика пространственного анализа. М., 2021. С. 370–371.

остатков минимизируют сумму квадратов остатков. А метод оценивания параметров регрессии a и b, который справляется с этой задачей, называют метод наименьших квадратов (остатков).

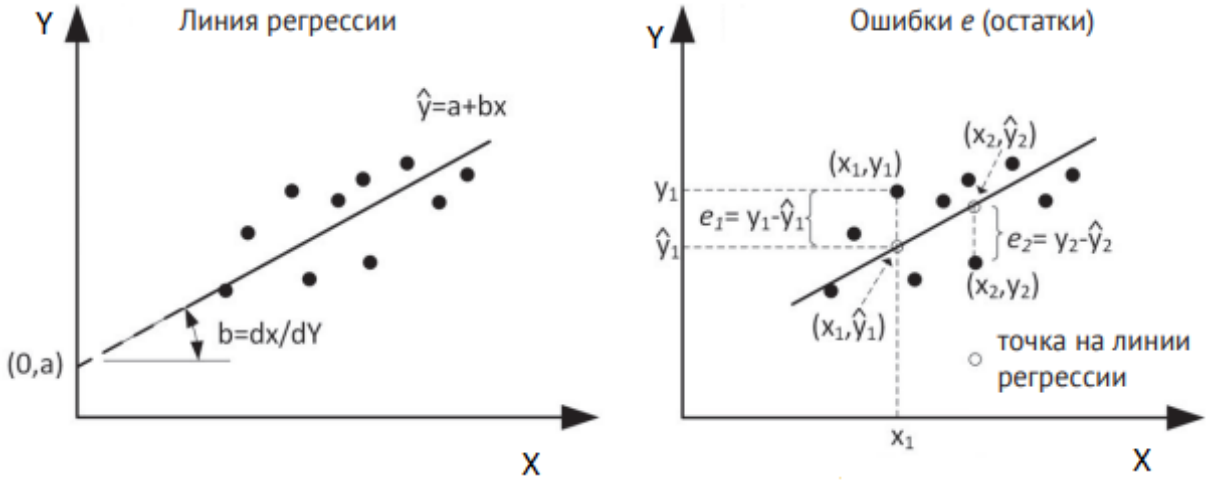

Рис. 4.1. Геометрическая интерпретация линейной парной регрессии[22]

Множественная регрессия представляет собой модель вида:

$$\hat{y} = f(x_1, x_2, \dots x_m). \qquad (4.5)$$

Для m независимых переменных и n наблюдений уравнение регрессии имеет вид:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_m x_m, \qquad (4.6)$$

где $\hat{y}$ – предсказанное (теоретическое, расчетное) значение зависимой переменной; b_1, b_2, … ,b_m – выборочные оценки коэффициентов регрессии; b_0 – выборочная оценка свободного коэффициента, определяет значение уравнения, когда независимые переменные или коэффициенты регрессии равны нулю; m – общее количество независимых переменных (регрессоров).

[22] *Грекусис Дж.* Методы и практика пространственного анализа. М., 2021. С. 371.

Эмпирическая вероятностная модель линейной парной регрессии имеет вид:

$$y = \hat{y} + e = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_m x_m + e, \qquad (4.7)$$

где y – фактическое значение зависимой переменной, e – остаток (ошибка) регрессии.

В матричном виде модель выражается, как показано ниже:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{nm} \end{bmatrix} \begin{bmatrix} b_0 \\ \vdots \\ b_m \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix},$$

$$Y = B \cdot X + e,$$

где Y – вектор с n наблюдениями зависимой переменной;

B – вектор оценок коэффициентов регрессии;

X – матрица данных с n наблюдениями для m независимых переменных;

e – вектор остатков (ошибок) регрессии.

Каждый коэффициент b_i можно интерпретировать как изменение среднего y при изменении на единицу переменной x_i при неизменности остальных независимых переменных. Знак коэффициента определяет направление влияния.

4.2. Метод наименьших квадратов и его предпосылки

Метод наименьших квадратов (МНК) позволяет получить такие оценки параметров a и b линейной парной регрессии, при которых сумма квадратов отклонений фактических значений регрессанта от предсказанных (теоретических) минимальна: $\sum_i (y_i - \hat{y})^2 \to \min$. Для нахождения минимума можно применить частные производные[23]:

$$S = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - a - bx_i)^2,$$

[23] *Картаев Ф.* Введение в эконометрику: учебник. М., 2019. С. 30.

$$\begin{cases} \dfrac{dS}{da} = -2\sum (y_i - a - bx_i) = 0, \\ \dfrac{dS}{db} = -2\sum x_i (y_i - a - bx_i) = 0, \end{cases}$$

$$\begin{cases} \sum y_i - na - b\sum x_i = 0, \\ \sum x_i y_i - a\sum x_i - b\sum x_i^2 = 0, \end{cases}$$

$$\begin{cases} \bar{y} - a - b\bar{x} = 0, \\ \overline{xy} - a\bar{x} - bx^2 = 0, \end{cases}$$

$$a = \bar{y} - b\bar{x},$$

$$b = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{\widehat{cov}(x,y)}{\widehat{var}(x)}. \qquad (4.8)$$

В случае линейной модели множественной регрессии МНК-оценки параметров регрессии можно определить векторно-матричным способом:

$$e = Y - B\cdot X,$$

$$Q = \sum {e_i}^2 = e^{'}\cdot e,$$

$$\frac{\partial Q}{\partial B} = -2X^{'}Y + 2\left(X^{'}X\right)B,$$

$$B = (X'X)^{-1} * X'Y, \qquad (4.9)$$

где Y – вектор значений зависимой переменной, B – вектор параметров регрессии, e – вектор остатков регрессии, X – матрица значений регрессоров.

Интерпретация. Возможность четкой экономической интерпретации коэффициента регрессии сделала линейное уравнение регрессии достаточно распространенным в эконометрических исследованиях. Выборочная оценка коэффициента регрессии b показывает среднее изменение результата с увеличением регрессора на одну единицу. Формально выборочная оценка свободного коэффициента a – это значение y при $x = 0$. Если x не имеет и не может иметь нуле-

вого значения, то такая трактовка свободного члена a не имеет смысла. Параметр a может не иметь экономического содержания. Попытки экономически интерпретировать его могут привести к абсурду, особенно при $a < 0$. Интерпретировать можно лишь знак при параметре a. Если $a > 0$, то относительное изменение регрессанта происходит медленнее, чем изменение регрессора.

Доказано, что надежность (эффективность, состоятельность, несмещенность) МНК-оценок регрессии a и b зависит от свойств случайного отклонения ε, которые называют ***предпосылками МНК:***

1) Математическое ожидание случайного отклонения равно нулю для всех наблюдений: $M(\varepsilon_i) = 0$.

2) Дисперсия случайных отклонений постоянна: $D(\varepsilon_i) = D(\varepsilon_j) = \sigma^2$. Выполнимость данной предпосылки называется *гомоскедастичностью остатков регрессии* (постоянством дисперсии отклонений). Невыполнимость данной предпосылки называется *гетероскедастичностью остатков регрессии* (непостоянством дисперсии отклонений), (см. рис. 4.2).

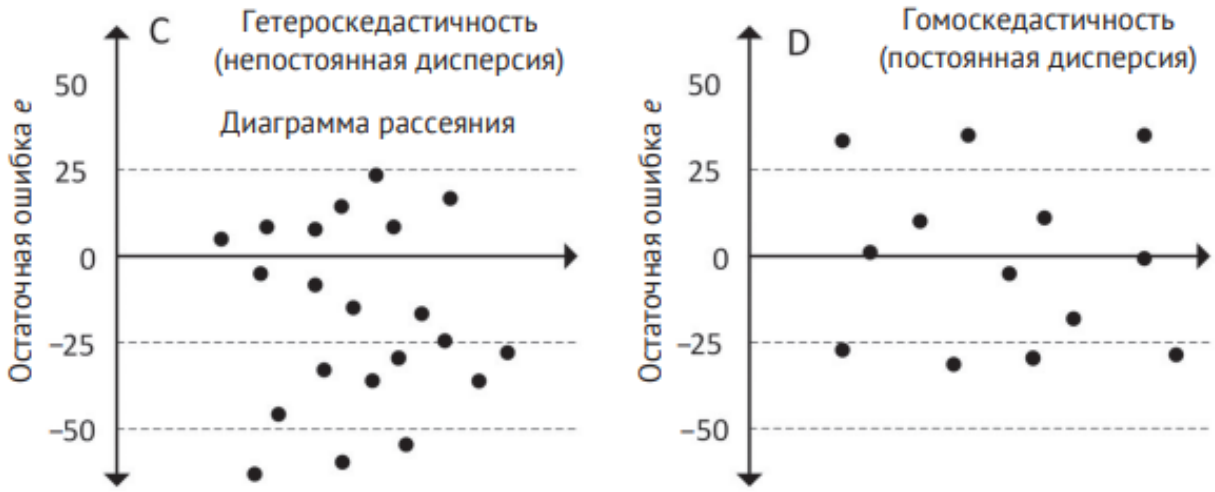

Рис. 4.2. Диаграммы гетероскедастичных и гомоскедастичных остатков регрессии[24]

[24] *Грекусис Дж.* Методы и практика пространственного анализа. М., 2021. С. 371.

3) Случайные отклонения ε_i и ε_j являются независимыми друг от друга для $i \neq j$:

$$cov(\varepsilon_i, \varepsilon_j) = \begin{cases} 0, & i \neq j \\ \sigma^2, & i = j. \end{cases}$$

Выполнимость этого условия называется отсутствием автокорреляции.

4) Случайное отклонение должно быть независимо от объясняющих переменных. Обычно это условие выполняется автоматически, если объясняющие переменные в данной модели не являются случайными. Кроме того, выполнимость данной предпосылки для эконометрических моделей не столь критична по сравнению с первыми тремя.

При выполнимости указанных предпосылок имеет место теорема Гаусса-Маркова: МНК-оценки параметров регрессии являются несмещенными, состоятельными и эффективными.

Помимо данных предпосылок метод линейной регрессии заключается в создании вероятностной модели на основе допущений о линейности взаимосвязи между y и x (проверка с помощью диаграммы рассеяния), нормальном распределении остатков с нулевым средним значением (проверка с помощью гистограммы и подобранной нормальной кривой, критерия Жарка-Бера), отсутствии мультиколлинеарности: переменные X_j должны быть независимыми друг от друга.

4.3. Коэффициент детерминации

Важный шаг в оценивании линейной регрессии - научиться измерять то, насколько полученное нами уравнение хорошо соответствует фактическим исходным данным. Выразим регрессант через остатки и предсказанные значения: $y_i = e_i + \hat{y}_i$. Теперь подсчитаем выборочную дисперсию переменной y_i, используя стандартные свойства выборочной дисперсии[25]:

$$S^2(y_i) = S^2(e_i + \hat{y}_i) = S^2(e_i) + S^2(\hat{y}_i) + 2\,cov(e_i, \hat{y}_i) = \\ = S^2(e_i) + S^2(\hat{y}_i) + 2 \cdot 0 = S^2(e_i) + S^2(\hat{y}_i),$$

[25] *Картаев Ф.* Введение в эконометрику: учебник. М., 2019. С. 32–33.

или, что тоже самое:

$$\frac{1}{n}\sum(y_i-\bar{y})^2=\frac{1}{n}\sum(y_i-\hat{y}_i)^2+\frac{1}{n}\sum(\hat{y}_i-\bar{y})^2. \qquad (4.10)$$

Слева в тождестве мы видим о***бщую сумму квадратов отклонений (TSS)***, справа соответственно ***остаточную (RSS)*** и ***регрессионную (ESS)***.

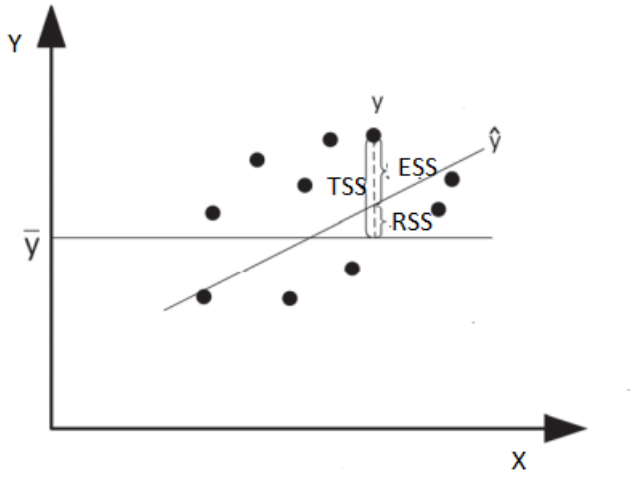

Рис. 4.3. Суммы квадратов отклонений

Если уравнение хорошо соответствует данным, то сумма квадратов остатков (RSS) должна быть маленькой (так как каждый из остатков регрессии близок к нулю). Если линия регрессии плохо соответствует фактическим данным, то сумма квадратов остатков будет большой. На этом сравнении основано использование ***коэффициента детерминации R^2***:

$$R^2=\frac{\sum(\hat{y}_i-\overline{y})^2}{\sum(y_i-\overline{y})^2}=1-\frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\overline{y})^2},$$
$$R^2=r^2{}_{yx}\,;0\le R^2\le 1. \qquad (4.11)$$

Коэффициент детерминации находится в диапазоне от 0 до 1, определяет долю разброса зависимой переменной y вокруг своего среднего значения, которая происходит под влиянием регрессоров x, и помогает оценить, насколько хорошо модель линейной регрессии

соответствует фактическим точкам данных. Нулевое значение указывает на то, что модель не объясняет никакой доли дисперсии и не может использоваться в дальнейшем. Значение 1 указывает на то, что модель объясняет всю изменчивость данных и линия регрессии идеально ложится на точки данных. Интерпретация коэффициента детерминации R-квадрат зависит от области и цели исследования. Когда требуется лишь определить наличие отношений между зависимыми и независимыми переменными, то может быть полезно даже низкое значение коэффициента детерминации R-квадрат.

Для линейной регрессии с константой верно, что общая сумма квадратов (TSS) равна сумме квадратов остатков (RSS) и объясненной сумме квадратов (ESS). Если константы в уравнении регрессии нет, то указанное равенство неверно, и R^2 не обязан лежать между нулем и единицей, интерпретировать стандартным образом его нельзя.

В случае линейной модели множественной регрессии, когда m – количество регрессоров x_j в модели увеличивается, то остаточная дисперсия уменьшается, и R^2 приблизится к 1 даже при слабой связи факторов с результатом. Добавление слишком большого числа переменных может привести к переобучению модели. Чтобы не допустить возможного преувеличения тесноты связи, используют ***скорректированный R^2***, который содержит поправку на число степеней свободы:

$$R^2{}_{adj} = 1 - \frac{\sum(y_i - \hat{y}_i)^2/(n-m-1)}{\sum(y_i - \bar{y})^2/(n-1)},$$

$$R^2{}_{adj} = 1 - (1 - R^2)\frac{n-1}{n-m-1}. \qquad (4.12)$$

R-квадрат предполагает, что каждая отдельная переменная объясняет какую-то часть дисперсии Y, тогда как скорректированный R-квадрат сообщает долю объясняемой дисперсии только для тех независимых переменных, которые влияют на Y. R-квадрат имеет тенденцию увеличиваться с добавлением каждой новой переменной, что вводит нас в заблуждение, что с каждым шагом мы создаем все луч-

шую модель. Фактически, добавляя все больше и больше переменных, мы приводим модель в состояние переобучения. Используя скорректированный R-квадрат, можно определить, какие переменные полезны, и включать в модель только те переменные, которые увеличивают скорректированный R-квадрат. Скорректированный R-квадрат корректируется вниз, по отношению к R-квадрату, и поэтому всегда меньше.

4.4. Проверка статистической значимости уравнения и коэффициентов регрессии

Тест Фишера проверяет нулевую гипотезу (H_0) ***о статистической незначимости уравнения регрессии*** и показателя тесноты связи (r_{yx}). Для линейной модели парной регрессии нулевая и альтернативная гипотезы имеют вид:

$$H_0 : \beta = 0,$$
$$H_1 : \beta \neq 0.$$

Для линейной модели множественной регрессии нулевая и альтернативная гипотезы имеют вид[26]:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_m = 0,$$
$$H_1 : \exists\, \beta_j \neq 0.$$

Критерием для проверки нулевой гипотезы является случайная величина, имеющая распределение Фишера:

$$F = \frac{R^2{}_{xy}}{1 - R^2{}_{xy}} \cdot \frac{n - m - 1}{m}.$$

$$F = \frac{\frac{\sum(\hat{y}_i - \bar{y})^2}{m}}{\frac{\sum(y_i - \hat{y})^2}{n - m - 1}},$$

[26] *Демидова О.А., Малахов Д.И.* Эконометрика: учебник и практикум для вузов. М., 2022. С. 120.

$$F > F_{\alpha,\nu_1=m,\nu_2=n-m-1} \Rightarrow H_1,$$
$$F < F_{\alpha,\nu_1=m,\nu_2=n-m-1} \Rightarrow H_0. \quad (4.13)$$

Отклонение нулевой гипотезы означает статистическую значимость уравнения регрессии в целом и подтверждает наличие статистической линейной взаимосвязи между регрессантом и регрессором (регрессорами). F-критерий проверяет совместную значимость всех коэффициентов, кроме свободного коэффициента, и также называется совместной F-статистикой. В общем случае F-статистика не особенно полезна, потому что довольно часто результаты являются статистически значимыми. Для оценки общего качества модели можно также использовать другие статистические критерии, такие как критерий Вальда и скорректированный R-квадрат.

Тест Стьюдента проверяет нулевую гипотезу (H_0) ***о статистической незначимости коэффициента регрессии***. Для модели линейной парной регрессии нулевые гипотезы тестов Стьюдента и Фишера одинаковы, поэтому можно ограничиться тестом Стьюдента[27]:

$$H_0 : \beta = 0,$$
$$H_1 : \beta \neq 0.$$

Критерием проверки нулевой гипотезы является случайная величина, имеющая распределение Стьюдента:

$$t_b = \frac{b}{m_b}; t_a = \frac{a}{m_a},$$

$$m_b = \sqrt{\frac{\sum \frac{(y-\hat{y})^2}{(n-2)}}{\sum (x-\overline{x})^2}} = \sqrt{\frac{S^2{}_{RSS}}{\sum (x-\bar{x})^2}} = \frac{S_{RSS}}{\sigma_x \sqrt{n}},$$

$$m_a = \sqrt{\frac{\sum (y-\hat{y})^2}{(n-2)} \cdot \frac{\sum x^2}{n \sum (x-\overline{x})^2}} = \sqrt{S^2{}_{RSS} \frac{\sum x^2}{n^2 \sigma^2{}_x}} = S_{RSS} \frac{\sqrt{\sum x^2}}{n\sigma_x},$$

[27] *Демидова О.А., Малахов Д.И.* Эконометрика: учебник и практикум для вузов. М., 2022. С. 100.

$$|t_b| > t_{\frac{\alpha}{2},n-2} \Rightarrow H_1,$$
$$|t_b| < t_{\frac{\alpha}{2},n-2} \Rightarrow H_0. \quad (4.14)$$

Отклонение нулевой гипотезы означает статистическую значимость коэффициента регрессии и подтверждает наличие статистической линейной взаимосвязи между регрессантом и регрессором.

Для модели линейной множественной регрессии нулевая гипотеза теста Стьюдента формулируется отдельно для каждого коэффициента регрессии.

Другим косвенным способом проверки статистической значимости коэффициента регрессии является расчет границ доверительного интервала:

$$\Delta b = t_{\frac{\alpha}{2},n-2} \cdot m_b,$$
$$b - \Delta b < \beta < b + \Delta b,$$
$$\Delta a = t_{\frac{\alpha}{2},n-2} \cdot m_a,$$
$$a - \Delta a < \alpha < a + \Delta a.$$

Если в границы доверительного интервала ноль не попадает, значит параметр регрессии является статистически значимым.

Тест Стьюдента используется для выбора независимых переменных для включения в модель. В случае, когда нулевая гипотеза о незначимости коэффициента регрессии не может быть отклонена, необходимо рассмотреть возможность удаления соответствующей независимой переменной из модели.

4.5. Стандартизованные остатки регрессии

Точность регрессионной модели в отношении предпосылок и допущений регрессии помогают проверить ***диаграммы остатков***. Наиболее часто используется диаграмма рассеяния, в которой остатки регрессии e откладываются по оси *Y*, а предсказанные значения $\hat{y}$ – по оси *X*. Когда остатки положительные, имеет место занижение прогно-

за, то есть вычисленное (прогнозируемое) значение ниже наблюдаемого. С другой стороны, когда остатки отрицательные, имеет место завышение прогноза, когда вычисленное (прогнозируемое) значение выше наблюдаемого. Если наблюдается систематическое занижение или завышение прогноза, это означает, что в модели присутствует систематическая ошибка. В идеале на диаграмме остатков не должно наблюдаться никаких закономерностей, а точки должны быть распределены случайно, иметь постоянную дисперсию и располагаться как можно ближе к оси *X*. Если на диаграмме остатков наблюдаются кластеры, выбросы или закономерности, это указывает на нарушение одной или нескольких предположений и допущений МНК.

С помощью графика распределения вероятностей остатков можно определить близость распределения остатков, а с помощью гистограммы остатков – асимметричность распределения и наличие выбросов.

Стандартизованные остатки регрессии могут быть получены путем деления остатков на оценку стандартного отклонения[28]. Стандартизированные остатки имеют среднее значение, равное нулю, и стандартное отклонение, равное 1. Диаграмма стандартизированных остатков помогает обнаружить большие расхождения в наборе данных, точки выбросов и влиятельные точки (см. рис. 4.4). ***Выбросы*** – это наблюдения, которые удалены от оси *X* более чем на 2 стандартных отклонения (в направлении оси *Y*). Обнаружив выброс, мы должны проверить, есть ли ошибка в наборе данных или значение, указывающее на необычное поведение. Влиятельной точкой является любая точка данных, которая существенно влияет на геометрию линии регрессии. Влиятельные точки должны выявляться и исключаться из анализа, потому что они существенно искажают геометрию линии регрессии и ухудшают точность моделей.

[28]*Грекусис Дж.* Методы и практика пространственного анализа. М., 2021. С. 397.

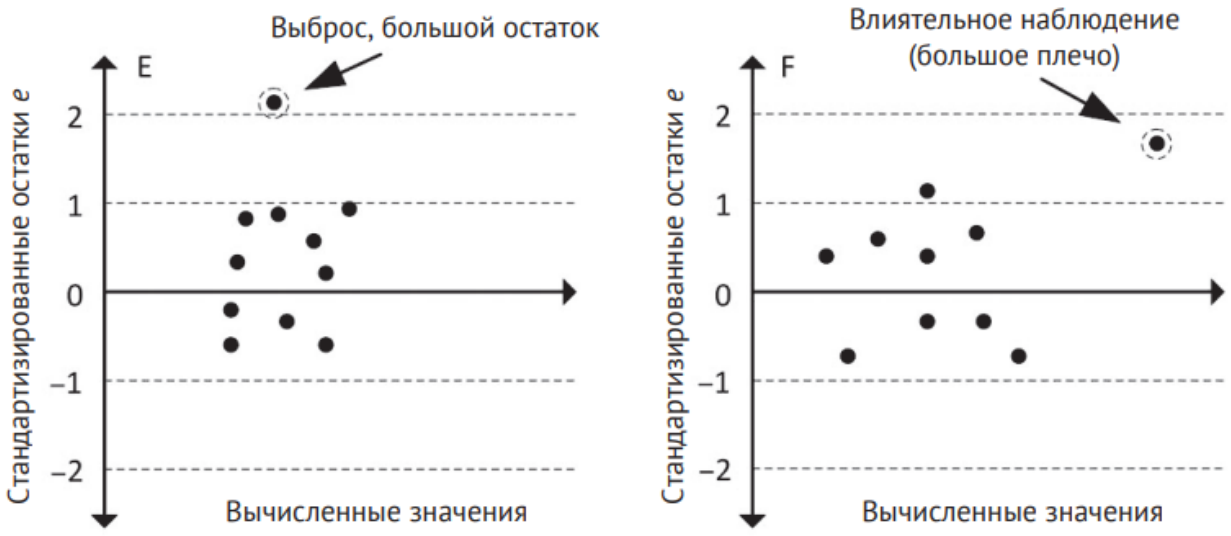

Рис. 4.4. Выбросы и влиятельные наблюдения[29]

Обобщение диагностических критериев, используемых в оценке качества линейной модели регрессии, выполнено в табл. 4.1.

Таблица 4.1

Непространственные статистические критерии, используемые в оценке качества линейной модели регрессии[30]

Критерий	Определяет	Проверяемая гипотеза (когда *p*-значение меньше уровня значимости, нулевая гипотеза отвергается)
F-статистика	Статистическую значимость модели	**Нулевая гипотеза:** значения коэффициентов регрессии, кроме свободного члена, равны нулю **Альтернативная гипотеза:** значения коэффициентов регрессии, кроме свободного члена, не равны нулю **Желаемый результат:** отклонение нулевой гипотезы

[29] *Грекусис Дж.* Методы и практика пространственного анализа. М., 2021. С. 402.
[30] *Грекусис Дж.* Методы и практика пространственного анализа. М., 2021. С. 475–477.

Продолжение табл. 4.1

Критерий	Определяет	Проверяемая гипотеза (когда *p*-значение меньше уровня значимости, нулевая гипотеза отвергается)
Совместный критерий Вальда	Статистическую значимость модели	**Нулевая гипотеза:** значения коэффициентов регрессии, кроме свободного члена, равны нулю **Альтернативная гипотеза:** значения коэффициентов регрессии, кроме свободного члена, не равны нулю **Желаемый результат:** отклонение нулевой гипотезы
t-статистика	Статистическую значимость отдельного коэффициента регрессии	**Нулевая гипотеза:** значение коэффициента регрессии равно нулю **Альтернативная гипотеза:** значение коэффициента регрессии не равно нулю **Желаемый результат:** отклонение нулевой гипотезы
Жарка-Бера	Ненормальность распределения остатков	**Нулевая гипотеза:** остатки распределены нормально **Альтернативная гипотеза:** распределение остатков не соответствует нормальному **Желаемый результат:** неотклонение нулевой гипотезы
Бройша-Пагана	Гетероскедастичность остатков	**Нулевая гипотеза:** остатки гомоскедастичны **Альтернативная гипотеза:** остатки гетероскедастичны **Желаемый результат:** неотклонение нулевой гипотезы
Коенкера	Гетероскедастичность остатков	**Нулевая гипотеза:** остатки гомоскедастичны **Альтернативная гипотеза:** остатки гетероскедастичны **Желаемый результат:** неотклонение нулевой гипотезы

Назначение. Линейная регрессия используется для: (а) выявления отношений между зависимой и независимой переменными, (б) оценки влияния (важности) независимой переменной на зависимую переменную, (в) создания прогнозирующей модели подгонкой линии регрессии под данные, (г) оценки правильности модели.

Глоссарий

F-критерий и уровень F-значимости – статистические критерии, которые используются для проверки статистической значимости линии регрессии, то есть насколько успешно модель объясняет значительную долю дисперсии наблюдаемых значений *Y*.

Влиятельная точка – это любая точка данных, которая существенно влияет на геометрию линии регрессии. Влиятельные точки должны выявляться и исключаться из анализа, потому что они существенно искажают геометрию линии регрессии и ухудшают точность моделей.

Диаграмма стандартизованных остатков – это диаграмма, в которой по оси абсцисс откладываются предсказанные значения зависимой переменной, а по оси ординат – стандартизованные остатки регрессии в единицах стандартного отклонения. Такая диаграмма помогает проверить соблюдение предположений МНК о нормальности и гомоскедастичности остатков. В идеале на диаграмме остатков не должно наблюдаться никаких закономерностей, а точки должны быть распределены случайно, иметь постоянную дисперсию и располагаться как можно ближе к оси *X*. Если в распределении точек наблюдаются кластеры, выбросы или закономерности, это указывает на нарушение одного или нескольких предположений.

Коэффициент детерминации (R-квадрат) – это процент дисперсии, объясняемой моделью. Он вычисляется как отношение объясненной суммы квадратов отклонений зависимой переменной (ESS) к общей сумме квадратов отклонений зависимой переменной (TSS). Изменяется в диапазоне от 0 до 1.

Множественная регрессия представляет собой модель, где теоретическое (среднее) значение зависимой переменной *Y* рассматривается как функция нескольких независимых переменных X_1, X_2,...X_m.

Обычный метод наименьших квадратов (Ordinary Least Squares, OLS) – это статистический метод оценки неизвестных параметров (коэффициентов) модели линейной регрессии. Регрессия наименьших квадратов подгоняет линию регрессии под имеющиеся данные путем минимизации суммы квадратов вертикальных расстояний от наблюдаемых точек до линии регрессии.

Остаток регрессии (несвязка) – это разница между наблюдаемым и предсказанным (вычисленным) значениями зависимой переменной.

Парная регрессия представляет собой модель, где теоретическое (среднее) значение зависимой переменной *Y* рассматривается как функция одной независимой переменной *X*.

Спецификация модели – формулирование вида модели, исходя из соответствующей теории связи между переменными.

Стандартизованные бета-коэффициенты – это коэффициенты, получающиеся при преобразовании переменных в регрессионной модели (как зависимых, так и независимых). Стандартизованные бета – коэффициенты выражаются в единицах стандартного отклонения, что исключает единицы измерения переменных. Используются для выбора независимых переменных *X*, наиболее важных для моделирования зависимой переменной *Y*.

Цель регрессионного анализа – оценка функциональной зависимости между независимыми переменными *X* и условным математическим ожиданием зависимой переменной *Y*.

Вопросы для самоконтроля

1. Чем вероятностная регрессионная модель отличается от уравнения регрессии?

2. Как определяется случайное отклонение модели?

3. В чем суть метода наименьших квадратов?

4. Каковы формулы расчета МНК-коэффициентов парного линейного уравнения регрессии?

5. Каковы предпосылки МНК?

6. Что такое коэффициент детерминации R-квадрат, и для чего он используется?

7. Что такое F-критерий и уровень F-значимости, и для каких целей они используются?

8. Что такое диаграмма стандартизированных остатков, и для чего она используется?

9. На каких предположениях основывается множественная линейная регрессия?

10. Что такое точки выброса и влиятельные точки?

ТЕМА 5. ПРОСТРАНСТВЕННАЯ ЭКОНОМЕТРИКА В ИЗМЕРЕНИИ ПРОСТРАНСТВЕННЫХ ЗАВИСИМОСТЕЙ. МАТРИЦЫ ВЕСОВ. БАЗА ДАННЫХ ГЛОБАЛЬНЫХ АДМИНИСТРАТИВНЫХ ОБЛАСТЕЙ GADM

В результате освоения темы обучающийся будет:

– ***знать*** сущность пространственной матрицы, способы ее построения;

– ***уметь*** выбирать тип матрицы весов под конкретные задачи исследования;

– ***владеть*** навыками построения матрицы весов.

Основные вопросы:

5.1. Пространственная эконометрика в измерении пространственных зависимостей.

5.2. Матрица весов: понятие и типы.

5.3. База данных глобальный административных областей GADM.

Ключевые слова: матрица весов, матрица сопряженности, граничная матрица, матрица расстояний, база GADM

5.1. Пространственная эконометрика в измерении пространственных зависимостей

Пространственный анализ или пространственная статистика включает в себя любые формальные методы, которые изучают объекты с использованием их топологических, геометрических или географических свойств. Пространственный анализ включает в себя множество методов, многие из которых используют различные аналитические подходы и применяются в различных областях. В более узком смысле пространственный анализ – это метод, применяемый при анализе географических данных. Для того чтобы подчеркнуть специфику объектов, расположенных не хаотично в пространстве, а в соответ-

ствии с определенными закономерностями, тенденциями, на начальном этапе анализа мы должны ответить на вопрос – существует ли пространственная зависимость в данных или нет? Если такая пространственная зависимость существует, то ее учитывают при построении регрессионных моделей в качестве дополнительного регрессора(-ов) и в таком случае применяют методы ***пространственной эконометрики*** (spatial econometrics), которая находится на стыке пространственного анализа и эконометрики и сосредотачивается на теоретических моделях, включающих пространственное взаимодействие между исследуемыми объектами.

Классификация методов пространственной эконометрики сложна из-за большого количества различных областей исследований, различных фундаментальных подходов, которые могут быть выбраны, и множества форм, которые могут принимать данные. Следовательно, мы остановимся на базовых методах, дающих концептуальное представление об алгоритме проведения пространственного анализа и построения пространственной модели.

Пространственная эконометрика как отдельное направление эконометрики обозначилось в конце XX века. Предпосылкой ее возникновения послужила новая экономическая география, которая получила свою популярность начиная со статьи П. Кругмана[31], будущего лауреата Нобелевской премии по экономике (2008 г.). Основная идея ***новой экономической географии*** состоит в том, что экономические показатели географических объектов зависят от их близости к другим географическим объектам. Мера близости регионов чаще всего измеряется с помощью географических расстояний, но в качестве такой меры может служить близость в торговле (например, объем товарооборота), миграционных потоках и т. д.

Для того чтобы учесть детальное влияние географических объектов друг на друга, в теоретическую модель необходимо вводить слишком много параметров (например, в виде фиктивных перемен-

[31] *Krugman P.R.* Increasing Returns and Economic Geography // The Journal of Political Economy. 1991. Vol. 99. № 3. P. 483–499.

ных: по одной переменной на каждую взаимосвязь). В таком случае их эмпирическая оценка становится невозможной. Следовательно, число оцениваемых параметров, характеризующих взаимное влияние географических объектов, стараются уменьшить. В то же время, если игнорировать переменные, характеризующие влияние других географических объектов, может возникнуть смещение в оценках параметров, и полученные выводы не будут адекватными. Многие идеи для построения пространственно-эконометрических моделей были позаимствованы из значительно более развитой теории временных рядов, только временные лаги заменялись на пространственные с помощью введения взвешивающей матрицы *W*, отражающей связи между географическими объектами[32].

Пространственные случайные процессы характеризуются следующими свойствами[33]:

1. Пространственная зависимость – наличие автокорреляции наблюдений. Выражается в невыполнении условия независимости остатков линейной регрессии. Устраняется с помощью пространственной регрессии.

2. Пространственная гетерогенность – нестационарность процессов, порождающих наблюдаемую переменную. Выражается в неэффективности постоянных коэффициентов линейной регрессии. Устраняется посредством географически взвешенной регрессии.

Для проверки статистических гипотез о наличии пространственных зависимостей мы будем использовать метод множителей Лагранжа. ***Метод множителей Лагранжа***, применяемый для решения задач математического программирования (в частности, линейного программирования) – метод нахождения условного экстремума

[32] *Демидова О.А.* Методы пространственной эконометрики и оценка эффективности государственных программ // Прикладная эконометрика. 2021. Т. 64. С. 108.

[33] *Самсонов Т.* Пространственная регрессия. Пространственная статистика и моделирование на языке R. URL: https://tsamsonov.github.io/r-spatstat-course/spreg.html (дата обращения: 10.12.2022).

функции *f(x)*, где $x \in R^2$, относительно m ограничений φ(x)=0, где i меняется от единицы до m.

Суть метода состоит в том, что функция Лагранжа *L(x)* составляется в виде линейной комбинации функции *f* и функций φ_i, взятых с коэффициентами, называемыми множителями Лагранжа – λ_i:

$$L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i \varphi_i(x),$$
$$\lambda = (\lambda_1 \dots, \lambda_m). \qquad (5.1)$$

Далее составляется система из $n+m$ уравнений, приравнивая к нулю частные производные функции Лагранжа *L(x,λ)* по x_j и λ_i. Если полученная система имеет решение параметров x_j' и λ_i, тогда точка x' может быть условным экстремумом, то есть решением исходной задачи. Заметим, что условие носит необходимый, но не достаточный характер.

Большим преимуществом этого метода является то, что он позволяет решать оптимизацию без явной параметризации в терминах ограничений. В результате метод множителей Лагранжа широко используется для решения сложных задач ограниченной оптимизации[34].

Базовыми моделями пространственной эконометрики являются модель пространственных лагов и модель пространственных ошибок (более подробно данные модели рассмотрены в отдельной теме).

SAR – модель с пространственным авторегрессионным лагом может быть представлена в виде:

$$y_i = \alpha + \rho \sum_{j=1}^{n} w_{ij} y_j + x_i \beta + \varepsilon_i, \text{где } \varepsilon_i \sim N(0, \sigma^2). \qquad (5.2)$$

Модель SAR используется в случаях, когда предполагается, что пространственные взаимодействия проявляются в том, что значения зависимой переменной для каждого региона зависят от значений той

[34] *Акулич И.Л.* Математическое программирование в примерах и задачах. М., 1986. 319 с.

же самой переменной в соседних регионах. Примерами подобных взаимодействий могут быть технологические «переливы» между соседними регионами[35].

SEM – модель с пространственным взаимодействием в ошибках может быть представлена в следующем виде[36]:

$$y = \alpha + x\beta + u,$$
$$u = \lambda W u + \varepsilon. \qquad (5.3)$$

Модель SEM применяется в случаях, когда пространственные взаимодействия между значениями зависимой переменной в соседних регионах маловероятны или несущественны, но при этом предполагается, что соседние регионы все же влияют друг на друга некоторым образом, который не отражается во включенных в модель регрессорах. Обе модели предполагают, что степень чувствительности регионов к влиянию соседей постоянна.

Для оценки моделей пространственной зависимости используют следующие методы[37]:

I. Для моделей SAR:
   – метод максимального правдоподобия (ML),
   – метод инструментальных переменных (IV).

II. Для моделей SEM:
   – метод максимального правдоподобия (ML),
   – обобщенный метод моментов (GMM).

В статистическом анализе ***метод максимального правдоподобия*** представляет собой метод оценки параметров вероятностного распределения с помощью максимизации функции правдоподобия.

[35] *Varlamova J., Larionova N.* Labor productivity in the digital era: a spatial-temporal analysis // International Journal of Technology. 2020. №11(6). P. 1191–1200.
[36] *Демидова О.А.* Методы пространственной эконометрики и оценка эффективности государственных программ // Прикладная эконометрика. 2021. Т. 64. С. 115.
[37] *Griffith D.A., Chun Y.* Spatial Regression Models. In: Huang B. (ed). Comprehensive geographic information systems. Elsevier, 2018. Vol. 3. P. 1–27.

Предположим, что случайный вектор $x = (X_1, X_2, ..., X_n)$ имеет плотность распределения $p(x; \theta)$, которая зависит от неизвестного параметра θ. Функцией правдоподобия называется случайная величина:

$$L = L(x; \Theta) = p(x; \Theta). \quad (5.4)$$

Оценкой максимального правдоподобия называется величина $\widehat{\Theta} = \widehat{\Theta}_{ML}$, которая максимизирует функцию правдоподобия L, то есть такая функция $\widehat{\Theta} = \widehat{\Theta}(x)$, что:

$$L\big(x; \widehat{\Theta}(x)\big) = \max_{\Theta} L(x; \Theta). \quad (5.5)$$

В регулярном случае необходимым условием максимума является уравнение $\frac{\partial L(x; \Theta)}{\partial \Theta} = 0$, которое называется уравнением правдоподобия.

Следует заметить, что если компоненты $X_1, X_2, ..., X_n$ вектора x независимы и одинаково распределены с плотностью $p(x; \theta)$, то функция правдоподобия есть произведение функцией правдоподобия каждой компоненты.

Для широкого круга задач оценки, полученные методом максимального правдоподобия являются состоятельными и асимптотически эффективными. Но в то же время они могут быть смещенными. Недостатком метода является необходимость знать распределение вектора x[38].

Метод инструментальных переменных используются в том случае, когда наблюдается сильная мультиколлинеарность между факторами, включенными в модель, или одна из существенных переменных в спецификации не имеет количественной оценки или не измеряется. Примером может служить талант работника, который наряду с образованием, возрастом, полом, является существенной

[38] *Магнус Я.Р., Катышев П.К., Пересецкий А.А.* Эконометрика. Начальный курс. М. 2021. С. 471.

переменной для моделирования заработной платы работника. Однако талант работника напрямую не измеряется и его достаточно сложно количественно оценить. Невключение существенной переменной в модель может привести к ошибке спецификации и неверным оценкам коэффициентов регрессии. Следовательно, необходимо подобрать переменную, которая выступала бы прокси-переменной для таланта работника.

5.2. Матрица весов: понятие и типы

Пространственная эконометрика интегрирует пространство и пространственные отношения непосредственно в математическое представление (площадь, расстояние, длина или близость). Как правило, эти пространственные отношения формально определены через значения, называемые ***пространственными весами***. Они используются для присвоения относительной значимости объектам.

Для исследования взаимосвязей между территориальными системами используются ***матрицы пространственных весов***, с помощью которых задается расстояние между объектами исследования.

Наиболее распространённой является ***матрица граничащих объектов***: строки матрицы содержат веса для объекта в пространстве, на который оказывают влияние соседние объекты. 0 – если объекты не имеют общую границу, 1 – если имеют общую границу[39].

$$W_{ij} = \begin{cases} 0, \text{ если } i=j \\ 1, \text{ если } j \text{ граничит с } i, \\ 0, \text{ если не граничит,} \end{cases}$$

где W_{ij} – элемент матрицы пространственных весов для регионов i и j.

[39] *Балаш В.А., Файзлиев А.Р.* Пространственная корреляция в статистических исследованиях // Вестник Саратовского государственного социально-экономического университета. 2008. № 4. С. 122–125.

Матрица является квадратной и ее главная диагональ состоит из нулей (для исключения влияние объекта на самого себя).

Таблица 5.2.1

Пример матрицы граничащих субъектов

	Кировская область	Нижегородская область	Оренбургская область	Пензенская область	Пермский край	Республика Башкортостан
Кировская область	0	1	0	0	1	0
Нижегородская область	1	0	0	0	0	0
Оренбургская область	0	0	0	0	0	1
Пензенская область	0	0	0	0	0	0
Пермский край	1	0	1	0	1	0
Республика Башкортостан	0	1	0	0	0	0

Граничная матрица представляет собой наиболее простой способ учета пространственных взаимосвязей. Согласно данной матрице на исследуемые объекты значительное влияние оказывают только те территории, которые граничат с ними. Данную матрицу можно использовать при однородности территорий по их размерам. Не рекомендуется использовать матрицу граничащих субъектов при значительных различиях регионов по площади[40].

При построении граничной матрицы можно использовать два подхода (см. рис. 5.1). В случае определения соседа по принципу

[40] *Боровиков В.* STATISTICA. Искусство анализа данных на компьютере: Для профессионалов. СПб., 2003. 688 с.

«ферзя», соседом считается каждый регион, имеющий хоть какую-нибудь граничащую точку с данным регионом. По принципу «ладьи» соседом считается регион, имеющий протяженную границу с данным регионом.

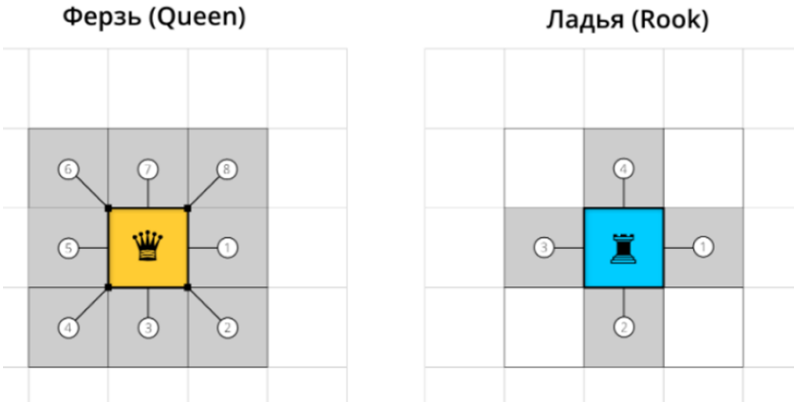

Рис. 5.1. Критерии определения смежности в граничной матрице[41]

Расширенной версией матрицы является **матрица «*k* ближайших соседей»:**

$$W_{ij}(K) = \begin{cases} 0, \text{ если } i=j, \\ 1, \text{ если } d_{ij} \le d_i(k), \\ 0, \text{ если если } d_{ij} > d_i(k), \end{cases}$$

где W_{ij} – элемент матрицы пространственных весов для регионов i и j, d_{ij} – расстояние от региона i до региона j, $d_i(k)$ – наибольшее из k наименьших расстояний[42].

В данной матрице понятие соседства учтено более широко и позволяет рассматривать в дальнейшем анализе связь не только соседних, но и близлежащих районов. В зависимости от значения

[41] *Лурье И.К., Самсонов Т.Е.* Информатика с основами геоинформатики. Часть 2. Основы геоинформатики. М., 2016. 200 с.

[42] *Балаш В.А., Файзлиев А.Р.* Пространственная корреляция в статистических исследованиях // Вестник Саратовского государственного социально-экономического университета. 2008. № 4. С. 122–125.

k каждый регион имеет большее или меньшее количество «ближайших» соседей.

Следующим видом является ***матрица расстояний*** (distance matrix), пространственные веса которой рассчитываются следующим образом[43]:

$$W_{ij} = \begin{cases} 0, \text{ если } i=j, \\ d_{ij}, \text{ если } i \neq j, \end{cases}$$

где W_{ij} – элемент матрицы пространственных весов для регионов i и j, d_{ij} – расстояние от региона i до региона j.

Матрица обратных расстояний – пространственная матрица, пространственные веса которой рассчитываются следующим образом:

$$W_{ij} = \begin{cases} 0, \text{ если } i=j, \\ 1/d_{ij}, \text{ если } i \neq j, \end{cases}$$

где W_{ij} – элемент матрицы пространственных весов для регионов i и j, d_{ij} – расстояние от региона i до региона j. Соответственно, чем дальше находится регион i от региона j, тем меньшему влиянию региона j от подвержен. Данный тип матрицы является одним из наиболее популярных. Кроме того, ряд исследователей использует взвешивающие матрицы, в которых по аналогии с теорией гравитации учитывается как географическая близость регионов, так и показатели мощности их экономик[44].

С помощью взвешивающих матриц можно создавать новые переменные – ***пространственные лаги***, показывающие средний уровень исследуемых показателей в соседних регионах.

---

[43] *Балаш В.А., Файзлиев А.Р.* Пространственная корреляция в статистических исследованиях // Вестник Саратовского государственного социально-экономического университета. 2008. № 4. С. 122–125.

[44] *Гурьянова Л.С., Холодный Г.А., Лукьянчикова А.С.* Методы и модели анализа пространственной кластеризации темпов социально-экономического развития регионов // Проблемы экономики. 2013. № 2. С. 242–250.

Выбор весовой матрицы остается дискуссионным вопросом. Ряд исследователей критиковали пространственные эконометрические модели именно за их чувствительность к выбору весовой матрицы[45], поскольку для различных типов весовых матриц можно получить разные оценки коэффициентов.

5.3. База данных глобальный административных областей GADM

На этапе построения картограмм и пространственных матриц возникает вопрос о загрузке пространственно организованных данных в специальную программную среду ГИС. При работе с географическими данными распространенным форматом файлов, содержащих географические координаты объектов являются так называемые «шейп-файлы» (shape-file). Шейп-файл представляет собой распространенный векторный формат хранения информации о геометрическом положении и определенных атрибутах географических объектов. Несмотря на свое название, в действительности шейп-файл – это не один, а набор из нескольких файлов с одинаковым именем, но разными расширениями. Обязательными при этом являются файлы с расширениями *.shp, .dbf* и *.shx*[46].

Одним из бесплатных (для некоммерческого использования!) источников является Global Administrative Areas (GADM) – Глобальная база данных административных областей. В этой базе данных доступны шейп-файлы трех разных уровней пространственного разрешения, например, на уровне страны, областей и районов. Текущая версия 4.1. включает 400276 административных районов мира. Дан-

[45] *Plümper T., Neumayer E.* Model specification in the analysis of spatial dependence // European Journal of Political Research. 2010. №49(3). P. 418–442. *Stakhovych S., Bijmolt T.H.* Specification of spatial models: A simulation study on weights matrices // Papers in Regional Science. 2009. № 88(2). P. 389–408.

[46] Создание картограмм при помощи R. URL: https://r-analytics.blogspot.com/2013/07/r.html (дата обращения: 10.12.2022).

ные находятся в свободном доступе для академического и другого некоммерческого использования[47].

Глоссарий

GADM – Глобальная база данных административных областей, содержащая шейп-файлы на уровне страны, областей и районов.

Граничная матрица – это пространственная матрица, элементы которой больше 0 для объектов, с которыми исследуемый объект имеет границу.

Матрица пространственных весов – это матрица, с помощью которой формализуется связь исследуемого объекта с другими объектами.

Матрица расстояний – пространственная матрица, веса которой определяются расстоянием между исследуемыми объектами.

Метод максимального правдоподобия позволяет найти оценки параметров распределения с помощью максимизации функции правдоподобия.

Метод множителей Лагранжа – метод нахождения условного экстремума функции относительно определенного количества ограничений.

Новая экономическая география – научное направление, основная идея которого состоит в том, что экономические показатели географических объектов зависят от их близости к другим географическим объектам.

Пространственная эконометрика (spatial econometrics) – научная дисциплина, которая находится на стыке пространственного анализа и эконометрики и сосредотачивается на теоретических моделях, включающих пространственное взаимодействие между исследуемыми объектами.

Пространственные лаги показывают средний уровень исследуемых показателей в соседних регионах.

[47] GADM. URL: https://gadm.org/ (дата обращения: 27.11.2022).

Вопросы для самоконтроля

1. В чем суть метода Лагранжа? В каких случаях его можно использовать?
2. В чем суть метода максимального правдоподобия? В каких случаях его можно использовать?
3. Что такое пространственная матрица? Для каких целей ее используют?
4. Какие бывают пространственные матрицы? Какие показатели могут быть использованы в качестве пространственных весов?
5. Каким образом строится граничная матрица? В чем ее достоинства и недостатки? В каких исследованиях ее можно использовать?
6. Каким образом строится матрица обратных расстояний? В чем ее достоинства и недостатки? В каких исследованиях ее можно использовать?
7. С чем связана критика пространственных матриц и пространственной эконометрики?

ТЕМА 6. ПРОСТРАНСТВЕННАЯ АВТОКОРРЕЛЯЦИЯ. ПРИМЕРЫ РАСЧЕТА ИНДЕКСОВ МОРАНА И ГИРИ И ПОСТРОЕНИЯ ДИАГРАММЫ МОРАНА ДЛЯ РЕГИОНОВ РОССИИ В НАУЧНЫХ СТАТЬЯХ

В результате освоения темы обучающийся будет:

– ***знать*** сущность пространственной автокорреляции, способы ее диагностики;

– ***уметь*** интерпретировать результаты расчетов индексов пространственной автокорреляции;

– ***владеть*** навыками диагностики пространственной автокорреляции.

Основные вопросы:

6.1. Понятие пространственной автокорреляции.

6.2. Диагностика пространственной автокорреляции.

6.3. Примеры диагностики пространственной автокорреляции для регионов России.

Ключевые слова: пространственная автокорреляция, глобальный индекс Морана, индекс Гири, диаграмма Морана, матрица Морана, индекс Гетиса-Орда

6.1. Понятие пространственной автокорреляции

Пространственная взаимосвязь статистически проявляется в наличии автокорреляции наблюдений. Она основывается на взаиморасположении объектов и их значениях. Если выборка включает в себя *n* пространственно организованных объектов, то пространственной автокорреляцией называется взаимосвязь между переменной, наблюдаемой у каждого из *n* объектов, и мерой их пространственной близости. При анализе исследователь исходит из того, что соседствующие территориальные системы связаны друг с другом намного больше, чем расположенные на значительном расстоянии. Понятие

пространственной автокорреляции является математическим отражением ***первого закона географии***: все связано со всем, но близкорасположенные объекты связаны сильнее. Коэффициенты пространственной автокорреляции позволяют сделать вывод о тесноте (силе) взаимосвязи между пространственно организованными объектами. Создание пространственной весовой матрицы позволяет количественно формализовать пространственную близость между объектами.

Пространственная автокорреляция используется, когда необходимо:

– оценить случайность распределения значений в пространстве;

– осуществить их пространственную кластеризацию;

– оценить концентрацию определённых показателей в пространстве;

– выявить тесноту пространственной взаимосвязи между территориальными группами.

Положительная пространственная автокорреляция выражается образованием групп с близкими показателями наблюдений (схожесть соседей). ***Отрицательная*** пространственная автокорреляция выражается формированием групп, которые существенно отличаются по своим характеристикам между собой (соседи не похожи друг на друга) (см. рис. 6.1).

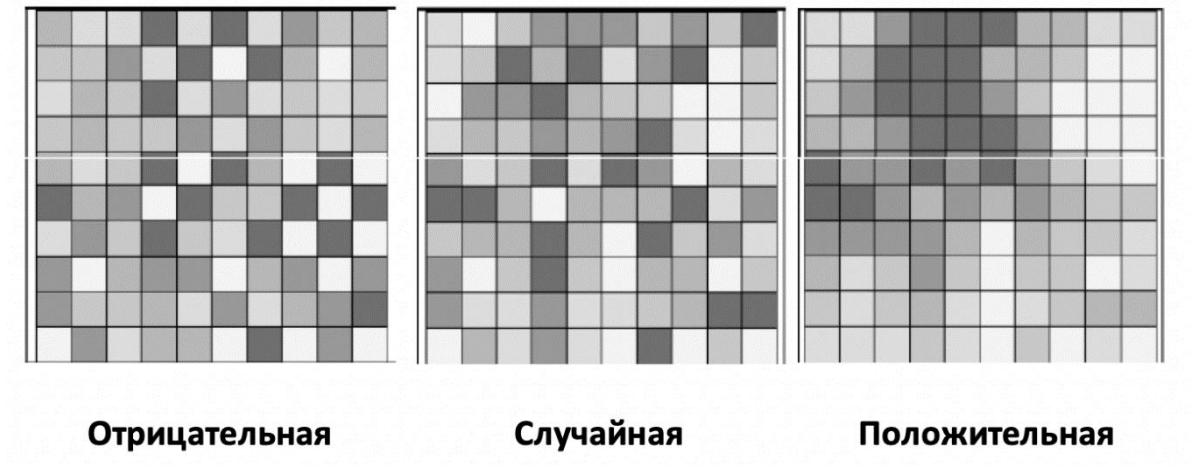

Рис. 6.1. Виды пространственной автокорреляции[48]

[48] *Окунев И.Ю.* Основы пространственного анализа. М., 2020. 255 с.

Пространственная автокорреляция выступает предметом острых дискуссий. Традиционный подход к построению регрессионных моделей не учитывает пространственной взаимосвязи между объектами. Пространственная эконометрика исходит из того, что учет пространственной автокорреляции важен с точки зрения правильности спецификации в модели регрессии. Для аналитика пространственных данных пространственная автокорреляция становится достаточно распространенным явлением, факт наличия которой необходимо учитывать при введении пространственного лага зависимой переменной, независимых переменных. Кроме того, пространственная взаимосвязь может наблюдаться в остатках модели, что также необходимо учитывать в спецификации.

В статистическом анализе может выделить несколько подходов, определяющих дальнейшие действия исследователя в случае присутствия пространственной автокорреляции в регрессионных моделях[49]:

1. Следует увеличить размер выборки до тех пор, пока не удастся устранить статистически значимую пространственную автокорреляцию. Представленный подход не гарантирует, что в выборочной совокупности пространственная автокорреляция проблема с пространственной автокорреляцией будет решена, но такой подход позволит ее уменьшить. Данный подход к уменьшению пространственной автокорреляции эффективен в том случае, если пространственная автокорреляция выступает результатом избыточности данных.

2. Методы ***фильтрации Гетиса*** позволяют изолировать пространственные и непространственные компоненты для каждой переменной. Алгоритм проведения фильтрации предполагает, что сначала пространство удаляют из каждой величины, а затем его возвращают обратно в регрессионную модель в качестве новой переменной, отвечающей за пространственные эффекты. Пространственная фильтрация

[49] Основы регрессивного анализа // ArcMap. URL: https://desktop.arcgis.com/ru/arcmap/latest/tools/spatial-statistics-toolbox/regression-analysis-basics.htm (дата обращения: 11.11.2022).

А. Гетиса эффективна в случае, когда данные распределены достаточно равномерно.

3. Пространственная эконометрика позволяет замоделировать пространственную автокорреляцию путем включения в модель пространственных лагов при соответствующих переменных.

6.2. Диагностика пространственной автокорреляции

Диагностика пространственной взаимосвязи предшествует этапу построения модели. Если предварительный анализ показывает наличие пространственной автокорреляции, то этот факт выступает в пользу построения пространственных эконометрических моделей. Следовательно, если на этапе диагностики пространственной автокорреляции мы получаем ее случайный характер, то дальнейшее моделирование может исключить построение моделей с пространственной автокорреляцией.

Мера пространственной автокорреляции вычисляется с помощью различных индексов. Одним из часто используемых является ***глобальный одномерный индекс Морана***. Индекс Морана оценивает статистическую взаимосвязь между значением показателя в данной территориальной единице и значениями исследуемого показателя в соседних территориальных единицах.

Индекс Морана рассчитывается по формуле[50]:

$$I = \frac{n}{W}\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}, \qquad (6.1)$$

где Y_i, Y_j – атрибутивные признаки объектов i и j соответственно; $\bar{x}$ – среднее значение признака по n объектам; w_{ij} – пространственный вес для пары объектов i и j; n – общее число объектов, W – сумма весов.

[50] *Anselin L.* An Introduction to Spatial Data Analysis. URL: http://geodacenter.github.io/workbook/5a_global_auto/lab5a.html#fn1 (date of access: 18.01.2023).

Дополнительно рассчитывается ожидаемое значение индекса Морана, с помощью которого делается вывод о статистической значимости индекса Морана [51]:

$$E(I) = -\frac{1}{n-1}, \tag{6.2}$$

где n – общее число объектов.

Значимость расчетного индекса Морана может оценить с помощью его сравнения с ожидаемым значением индекса *E(I)* и его стандартным отклонением. Для такой оценки используется Z-тест Фишера:

$$Z = \frac{I - E(I)}{\sqrt{E(I^2) - E(I)^2}}. \tag{6.3}$$

Важно заметить, что индекс Морана по своей сути напоминает линейный коэффициент корреляции Пирсона (формула 6.4), в котором оценивается взаимосвязь между всеми парами значений переменных *X* и *Y*:

$$r_{xy} = \frac{\sum(x_i - \overline{x}) * (y_i - \bar{y})}{\sqrt{\sum(x_i - \overline{x})^2} * \sqrt{\sum(y_i - \bar{y})^2}}, \tag{6.4}$$

где x_i – значения переменной *X*; y_i – значения переменной *Y*; $\bar{x}$ – среднее арифметическое для переменной *X*; $\bar{y}$ – среднее арифметическое для переменной *Y*.

При расчете индекса Морана определяется взаимосвязь, но на уровне соседства между *i*-той и *j*-той территориальными единицами. Степень тесноты или силы соседства задается с помощью специального веса w_{ij}, который представлен в числителе формулы ин-

[51] *Anselin L.* An Introduction to Spatial Data Analysis. URL: http://geodacenter.github.io/workbook/5a_global_auto/lab5a.html#fn1 (date of access: 18.01.2023).

декса Морана (формула 6.1). Следовательно, если территориальные единицы пространственно не связаны друг с другом, значит, для них $w_{ij} = 0$, и эти пары не участвуют в вычислении индекса Морана.

Индекс Морана, рассчитанный для нормально распределенных величин, принимает значения в интервале от -1 до 1, при этом:

+1 характеризует прямую зависимость – группировку схожих (низких или высоких) значений;

0 означает абсолютно случайное распределение, то есть взаимосвязь между значением показателя в данной территориальной единице и значениями в территориях-соседях отсутствует;

-1 означает детерминированную обратную зависимость – чередование низких и высоких значений, напоминающее шахматную доску (см. рис. 6.1)[52].

При использовании стандартизированной по строкам весовой матрицы W = n, происходит модификация уравнения (6.1) следующим образом:

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}. \qquad (6.5)$$

Пространственная диаграмма рассеяния Морана[53] показывает линейную аппроксимацию облака точек, когда по оси абсцисс откладывается значение признака, а по оси ординат пространственный лаг признака – значение вектора *Wy*, то есть взвешенное среднее значение признака в соседних регионах (см. рис. 6.2). Наклон линии соответствует индексу Морана *I*, а его значение (0,103) указано в верхней части графика.

---

[52] *Демидова О.А., Камалова Э.* Пространственно-эконометрическое моделирование экономического роста российских регионов: имеют ли значение институты? // Экономическая политика. 2021. Т. 16. № 2. С. 34–59.

[53] *Anselin L.* An Introduction to Spatial Data Analysis. URL: http://geodacenter.github.io/workbook/5a_global_auto/lab5a.html#fn1 (date of access: 18.01.2023).

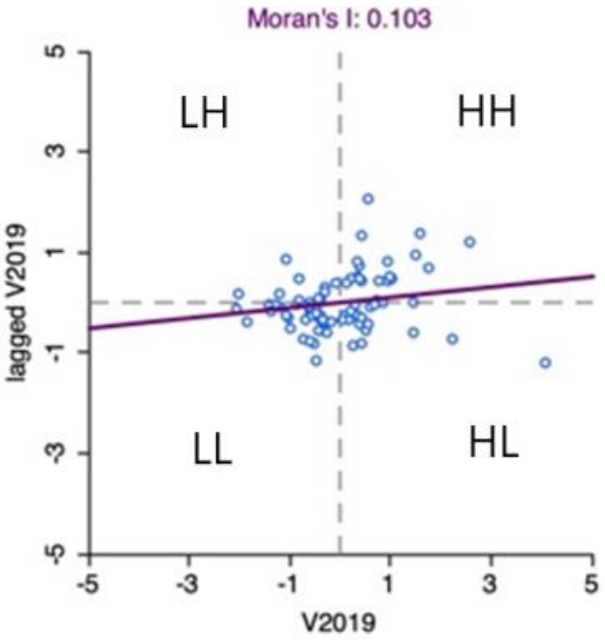

Рис. 6.2. Диаграмма рассеяния Морана

На основе диаграммы Морана составляется ***матрица Морана***, отражающая тип пространственной близости между различными объектами (см. табл. 6.1) [54].

Таблица 6.1

Матрица Морана

Квадрант	Описание
High-High	Регионы с высокими значениями анализируемого показателя окружены регионами также с относительно высокими значениями показателя
Low-High	Регионы с низкими значениями анализируемого показателя окружены регионами также с относительно высокими значениями показателя

[54] *Наумов И.В., Отмахова Ю.С., Красных С.С.* Методологический подход к моделированию и прогнозированию воздействия пространственной неоднородности процессов распространения COVID-19 на экономическое развитие регионов России // Компьютерные исследования и моделирование. 2021. Т. 13. № 3. С. 629–648.

Продолжение табл. 6.1

Квадрант	Описание
Low-Low	Регионы с низкими значениями анализируемого показателя окружены регионами также с относительно низкими значениями показателя
High-Low	Регионы с высокими значениями анализируемого показателя окружены регионами также с относительно низкими значениями показателя

Двумерный индекс Морана используется, если пространственное распределение по одной переменной коррелирует с пространственным распределением по другой переменной. Например, если выдвигается гипотеза, что пространственная автокорреляция числа врачей связана с пространственной автокорреляцией по заболеваемости населения.

Аналогично, гипотеза о случайном расположении регионов может быть проверена с помощью индекса Гири, который вычисляется по формуле:

$$C(Y) = \frac{(n-1)}{2n} \frac{\sum_{i,j=1}^{n} w_{ij}\left(Y_i - Y_j\right)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}, \qquad (6.6)$$

и принимает значения $0 < C(Y) < 2$. Если C значим и меньше 1, то имеет место положительная пространственная зависимость, а если значим и больше 1, то имеет место отрицательная пространственная зависимость[55].

[55] *Демидова О.А.* Методы пространственной эконометрики и оценка эффективности государственных программ // Прикладная эконометрика. 2021. Т. 64. С. 112–113.

Пространственная автокорреляция ***по методологии А. Гетиса и Дж. Орда*** используется, когда данные распределены достаточно равномерно. Индекс Гетиса-Орда может быть представлен в виде[56]:

$$G = \frac{\sum_{i,j=1}^{n} w_{ij} Y_i Y_j}{\sum_{i,j=1}^{n} Y_i Y_j}. \quad (6.7)$$

Ожидаемое среднее значение индекса автокорреляции будет вычисляться по формуле:

$$E(G) = \frac{\sum_{i,j=1}^{n} w_{ij}}{n(n-1)}, \quad (6.8)$$

где Y_i, Y_j – атрибутивные признаки объектов; w_{ij} – пространственный вес для пары объектов i и j; n – общее число объектов.

При сравнении индекса Гетиса-Орда и среднего значения индекса автокорреляции можно получить следующие типы кластеризации (см. табл. 6.2).

Таблица 6.2

Типы пространственной кластеризации по методике Гетиса-Орда

$G > E(G)$	Наблюдается пространственная кластеризация объектов с высокими значениями
$G < E(G)$	Наблюдается пространственная кластеризация объектов с низкими значениями

Методику Гетиса-Орда не рекомендуется применять, если наблюдается неоднородность распределения показателей по территориальным системам. В условиях высокой поляризованности данных

[56] *Демидова О.А.* Методы пространственной эконометрики и оценка эффективности государственных программ // Прикладная эконометрика. 2021. Т. 64. С. 112–113.

метод дает ложные результаты и рекомендуется использовать индекс Морана[57].

6.3. Примеры диагностики пространственной автокорреляции для регионов России

В работе «Пространственная корреляция в статистических исследованиях» В.А. Балаш и А.Р. Файзлиев[58] анализируют пространственную зависимость между торговыми площадями города Саратова. Выборку исследования составили 21 микрорайон г. Саратова, по торговым площадям в 2007 г. В результате анализа пространственных диаграмм рассеяния Морана, которые были построены с помощью граничной матрицы и матрицы обратных расстояний была установлена положительная пространственная корреляция между торговыми площадями города, то есть площадки с низким значением признака окружены площадками также с низкими значениями. Была установлена одна площадка с отрицательной автокорреляцией – она имела высокие значения при низких значениях в соседних территориях. Пример исследования показывает, что результаты диагностики пространственной автокорреляции могут меняться в зависимости от типа пространственной взвешивающей матрицы.

Другим примером диагностики пространственной автокорреляции является статья «Пространственные взаимодействия: оценка на основе локального и глобального индексов Морана» Ю.В. Павлова и Е.Н. Королёвой[59]. В работе оценена сила взаимосвязи и взаимовлияния между территориальными единицами на основе расчетов глобальных и локальных индексов Морана. В исследовании в качестве

---

[57] *Боровиков В.* STATISTICA. Искусство анализа данных на компьютере: для профессионалов. СПб., 2003. 688 с.

[58] *Балаш В.А., Файзлиев А.Р.* Пространственная корреляция в статистических исследованиях // Математические и инструментальные методы экономики. 2008. № 4 (23). С. 122–125.

[59] *Павлов Ю.В., Королёва Е.Н.* Пространственные взаимодействия: оценка на основе локального и глобального индексов Морана // Пространственная экономика. 2014. № 3. С. 95–109.

объектов были взяты 37 муниципальных образований 1-го уровня – 10 городских округов и 27 муниципальных районов Самарской области.

На основе глобального индекса Морана построена пространственная диаграмма рассеяния, выявлены четыре территориальных кластера в пределах Самарской области, далее с помощью локального индекса были определены шесть подкластеров. «Достоинством использования глобального индекса Морана в работе явилось то, что он позволил выявить три ядра. Недостатком является чрезмерная общность полученных результатов, так как степень влияния ядер на различные территории в пределах кластера неодинакова. Локальный индекс Морана позволил выявить подкластеры, образованные территориями со схожими значениями автокорреляции. Данный индекс рассчитывается по силе взаимовлияния между двумя конкретными территориями»[60].

Еще одним примером измерения пространственной автокорреляции является статья «Цифровизация промышленного производства в регионах России: пространственные взаимосвязи» И.В. Наумова, Ю.В. Дубровской, Е.В. Козоноговой[61]. Авторы построили и использовали миграционную матрицу пространственных весов, глобальные и локальные индексы Морана для исследования пространственной неоднородности цифровой трансформации промышленности по регионам России; составили матрицу локальных индексов автокорреляции Л. Анселина для измерения межрегиональных взаимосвязей в процессах использования цифровых технологий производственными предприятиями. Пример статьи показывает возможность создания весовой матрицы, исходя из потребностей исследовательской гипоте-

[60] *Павлов Ю.В., Королёва Е.Н.* Пространственные взаимодействия: оценка на основе локального и глобального индексов Морана // Пространственная экономика. 2014. № 3. С. 95–109.

[61] *Наумов И.В., Дубровская Ю.В., Козоногова Е.В.* Цифровизация промышленного производства в регионах России. Пространственные взаимосвязи // Экономика региона. 2020. Т. 16. Вып. 3. С. 896–910.

зы не учитывающей внутрирегиональные потоки, а также измерения корреляции между территориальными системами.

В одной из немногих работ для муниципалитетов “The impact of digitalization on the demand for labor in the context of working specialties: Spatial analysis” Е.В. Дубровская и Е.В. Козоногова[62] на примере районов Пермского края проверяют гипотезу о значимости расположения и соседства территорий для спроса на рабочую силу в условиях цифровизации экономики. Авторы обнаружили высокую пространственную неоднородность уровня регистрируемой безработицы, центры локализации и развития трудовых ресурсов, показали, что периферийные муниципальные образования с высоким уровнем регистрируемой безработицы имеют самый высокий среди других групп профессий коэффициент локализации востребованных специалистов в группе «Информационные технологии».

Использование индексов Морана для оценки различных параметров позволяет выявить варианты территориальных кластеров и подкластеров. Это даёт возможность выявить пространственные эффекты и принять оптимальные управленческие решения в зависимости от поставленных целей.

Глоссарий

Глобальный одномерный индекс Морана – индекс, направленный на выявление пространственной автокорреляции между значениями показателя в данной территориальной единице и значениями показателями в соседних территориальных единицах.

Двумерный индекс Морана используется, если пространственное распределение по одной переменной коррелирует с пространственным распределением по другой переменной.

[62] *Dubrovskaya J.V., Kozonogova E.V.* The impact of digitalization on the demand for labor in the context of working specialties: Spatial analysis // Вестник СПбГУ. Экономика. 2021. № 37(3). С. 395–412.

Матрица Морана отражает тип пространственной близости между различными объектами.

Отрицательная пространственная автокорреляция выражается формированием групп, которые существенно отличаются по своим характеристикам между собой (соседи не похожи друг на друга).

Положительная пространственная автокорреляция выражается образованием групп с близкими показателями наблюдений (схожесть соседей).

Пространственная автокорреляция – это наличие взаимосвязи между наблюдениями, которая основывается на взаиморасположении объектов и их значениях.

Пространственная диаграмма рассеяния Морана показывает линейную аппроксимацию облака точек, когда по оси абсцисс откладывается значение признака, а по оси ординат пространственный лаг, то есть взвешенное среднее значение признака в соседних регионах.

Вопросы для самоконтроля

1. Почему диагностика пространственной автокорреляции предшествует построению регрессионных моделей?
2. Какие типы пространственной автокорреляции можно выделить?
3. Какими методами можно диагностировать пространственную автокорреляцию?
4. Каким образом определяется попадание наблюдение в соответствующий квадрант матрицы Морана?
5. Определите, каким образом была проверена значимость индекса Морана в работе «Пространственная корреляция в статистических исследованиях» (В.А. Балаш, А.Р. Файзлиев).
6. Какой тип матрицы был использован в работе «Пространственные взаимодействия: оценка на основе локального и глобального индексов Морана» Ю.В. Павлова и Е.Н. Королёвой?

ТЕМА 7. СТАТИЧЕСКИЕ ПРОСТРАНСТВЕННЫЕ ЭКОНОМЕТРИЧЕСКИЕ МОДЕЛИ SAR, SDM, SEM. ПРЯМЫЕ И КОСВЕННЫЕ ЭФФЕКТЫ

В результате изучения данной темы обучающийся будет:

– ***знать*** определение и типы пространственно-эконометрических моделей; проблемы метода наименьших квадратов в случае пространственной автокорреляции и пространственной неоднородности; взаимосвязь между моделями пространственной регрессии; спецификацию и назначение модели пространственного лага; спецификацию и назначение модели пространственной ошибки;

– ***уметь*** выбрать тип пространственно-эконометрической модели и находить ее оценки; определять качество подгонки пространственно-эконометрических моделей; определять и интерпретировать прямые, косвенные и общие эффекты;

– ***владеть*** навыками выбора пространственно-эконометрических моделей, интерпретации пространственно-эконометрических моделей.

Основные вопросы:

7.1. Статические и динамические пространственные эконометрические модели. Типы статических пространственных эконометрических моделей.

7.2. Модель пространственного лага.

7.3. Модель пространственной ошибки.

7.4. Прямые и косвенные эффекты.

7.5. Выбор спецификации пространственно-эконометрической модели.

Ключевые слова: пространственная автокорреляция, пространственная неоднородность, пространственный лаг, весовая матрица, пространственно-эконометрическая модель, прямой эффект, косвенный эффект, общий эффект, коллекция множителей Лагранжа, коэф-

фициент при пространственном лаге, коэффициент при пространственной ошибке.

7.1. Статические и динамические пространственно-эконометрические модели. Типы статических пространственно-эконометрических моделей

Пространственная зависимость (пространственная автокорреляция) подразумевает функциональную связь между происходящим в некотором месте и в его окрестностях. ***Пространственная неоднородность (пространственная структура)*** возникает при отсутствии однородности в пространстве. Модели пространственной эконометрики, называемые также пространственно-эконометрическими моделями или моделями пространственной регрессии, оценивают пространственную зависимость и учитывают пространственную неоднородность для выявления взаимосвязей между зависимыми и независимыми переменными с привязкой к географическому местоположению. Пространственная эконометрика применяется в основном в региональных исследованиях.

В основе пространственно-эконометрических моделей лежит достаточно простая идея: при моделировании показателей территорий надо учитывать не только влияние других факторов на этих территориях, но и значения этих же показателей на других территориях. Однако введение отдельного параметра для учета влияния каждой из территорий уменьшает число степеней свободы в общей модели. Поэтому в пространственно-эконометрических моделях используют введение ***взвешивающей матрицы W*** для сокращения количества оцениваемых параметров. Число параметров, отражающих влияние других регионов, может быть сокращено до одного – коэффициента пространственной автокорреляции (по аналогии с коэффициентом автокорреляции во временных рядах). Если этот коэффициент оказывается значимым и положительным (отрицательным), то делают вывод о существовании положительных (отрицательных) пространственных

эффектов, то есть какое-либо изменение, произошедшее в одном регионе, приведет к аналогичному (противоположному) по действию изменению в соседнем регионе (если используется граничная матрица). Идеи многих других пространственных моделей также почерпнуты из теории временных рядов, только в пространственных моделях временные лаги Y_{t-1}, X_{t-1}, e_{t-1} заменяются на соответствующие пространственные лаги WY, WX, We , где W – взвешивающая матрица, отражающая влияние всех остальных регионов.

При наличии пространственной автокорреляции наблюдения могут быть пространственно сгруппированы или рассредоточены. В данных могут существовать географические тенденции, которые нарушают предположение о независимости наблюдений. При использовании классической регрессии методом наименьших квадратов и наличии пространственной автокорреляции:

– коэффициенты корреляции и коэффициент детерминации (R-квадрат) оказываются больше, чем есть на самом деле (смещены вверх), что завышает представление о тесноте взаимосвязи;

– стандартные ошибки и дисперсии коэффициентов регрессии занижены, поэтому p-значения в тестах Стьюдента и Фишера занижены, регрессоры и уравнение в целом могут быть ошибочно признаны статистически значимыми и ошибочно сделан вывод о взаимосвязи, которая в действительности отсутствует.

Пространственная неоднородность нарушает предположение о существовании единой линейной зависимости для всего набора данных, потому что взаимосвязь меняется в зависимости от местоположения[63]. Поэтому нарушается предположение метода наименьших квадратов о независимости наблюдений и нормальном распределении остатков. Пространственная неоднородность вызывает ошибки в оценке уровня значимости; неоптимальные прогнозы; признание существования отношений, которые отсутствуют в действительности; и ложное представление о том, что результаты применяются ко всему

[63] *LeSage J.P.* An Introduction to Spatial Econometrics. Revue D Économie Industrielle. 2008. 123(123). P. 19–44.

набору данных (глобально), хотя модель может быть точной только для определенных географических регионов.

Три различных типа эффектов пространственного взаимодействия могут объяснить, почему наблюдение, связанное с конкретным местом, может зависеть от наблюдений в других местах. Во-первых, это эффекты эндогенного взаимодействия, когда зависимая переменная отдельной территории А зависит от зависимой переменной других территорий, среди которых, скажем, территория Б, и наоборот[64]:

Зависимая переменная Y территории A ↔
Зависимая переменная Y территории B

Географической единицей территории могут быть почтовые индексы, города, муниципалитеты, регионы, округа, страны. Эффекты эндогенного взаимодействия обычно рассматриваются как формальная спецификация равновесного результата процесса пространственного или социального взаимодействия, в котором значение зависимой переменной для одного агента определяется совместно со значением соседних агентов. Например, эффекты эндогенного взаимодействия теоретически согласуются с ситуацией, когда налогообложение и расходы на общественные услуги на одной территории взаимодействуют с налогообложением и расходами на общественные услуги в соседних территориях.

Во-вторых, эффекты экзогенного взаимодействия, где зависимая переменная отдельной территории А зависит от независимых объясняющих переменных других территорий, в том числе, территории Б:

Зависимая переменная Y территории A ↔
Независимая переменная X территории B

Например, как в теоретической, так и в эмпирической литературе по экономическому росту и конвергенции между странами/регионами

[64] *Elhorst J.P.* Spatial Econometrics: from Cross-Sectional Data to Spatial Panels. 2014. P. 7–8.

переменная экономического роста считается зависящей не только от исходного уровня дохода и норм сбережений, прироста населения, технологических изменений и амортизации в собственной экономике, но и от этих переменных в соседних странах/регионах.

В-третьих, третий тип эффектов взаимодействия – это те, которые относятся к ошибкам:

Член ошибки территории A ↔ Член ошибки территории B

Эффекты взаимодействия между членами ошибки согласуются с ситуацией, когда имеются пространственно автокоррелированные и не включенные в модель существенные регрессоры, или ненаблюдаемые шоки. Например, непредвиденные изменения налогово-бюджетной политики в результате действий политиков, стремящихся к получению ренты[65].

Для оценки влияния изложенных эффектов пространственного взаимодействия в регрессионном анализе было предложено несколько моделей. Пространственная зависимость добавляется к регрессии путем включения пространственного лага и пространственной ошибки. Модель со всеми эффектами взаимодействия называется общей вложенной пространственной моделью ***(General nesting spatial model, GNS)***[66]:

$$Y = \alpha + \rho WY + X\beta + WX\theta + u,$$
$$u = \lambda Wu + \varepsilon, \qquad (7.1)$$

где WY – эффекты эндогенного взаимодействия между зависимой переменной; WX – эффекты экзогенного взаимодействия между независимыми переменными и зависимой переменной; Wu – эффекты взаимодействия между членами ошибок на разных территориях; ρ – про-

[65] *LeSage J.P.* An Introduction to Spatial Econometrics. Revue D Économie Industrielle. 2008. 123(123). P. 19–44.

[66] *Elhorst J. P.* Spatial Econometrics: from Cross-Sectional Data to Spatial Panels. 2014. P. 9.

странственный авторегрессионный коэффициент; θ как и β, представляет вектор $k \times 1$ неизвестных параметров для оценивания; W – неотрицательная матрица $n \times n$ описывающая пространственную конфигурацию территорий в выборке данных;

– модель с пространственным лагом зависимой переменной и пространственной ошибкой ***(SAC)***:

$$Y = \alpha + \rho WY + X\beta + u, u = \lambda Wu + \varepsilon, \quad (7.2)$$

– модель Дарбина с пространственными лагами объясняющих переменных и пространственным лагом зависимой переменной ***(Spatial Durbin model, SDM)***:

$$Y = \alpha + \rho WY + X\beta + WX\theta + \varepsilon. \quad (7.3)$$

– модель Дарбина с пространственными лагами объясняющих переменных и пространственной ошибкой ***(Spatial Durbin Error model, SDEM)***:

$$\begin{gathered} Y = \alpha + X\beta + WX\theta + u, \\ u = \lambda Wu + \varepsilon. \end{gathered} \quad (7.4)$$

– пространственная авторегрессионная модель ***(Spatial Autoregressive model, SAR)***:

$$Y = \alpha + \rho WY + X\beta + \varepsilon. \quad (7.5)$$

– модель с пространственными лагами объясняющих переменных ***(Spatial Lag of X model, SLX)***:

$$Y = \alpha + X\beta + WX\theta + \varepsilon. \quad (7.6)$$

– модель пространственной ошибки ***(Spatial Error model, SEM)***:

$$\begin{gathered} Y = \alpha + X\beta + u, \\ u = \lambda Wu + \varepsilon. \end{gathered} \quad (7.7)$$

Рисунок 7.1 систематизирует пространственные эконометрические модели, среди которых модель МНК – справа, и модель GNS –

слева, использованы спецификации в матричном виде, составленные Д. Элхорстом. Каждая модель справа от модели GNS может быть получена из этой модели путем наложения ограничений на один или несколько ее параметров.

Наиболее привлекательными для изучения теоретических эконометрических проблем (в частности, условий стационарности для параметров и пространственной матрицы) являются модели SAR, SEM, SAC. Наименее проблематичными в оценивании стандартными методами являются пространственная эконометрическая модель с экзогенными эффектами – SLX, и модель с экзогенными эффектами взаимодействия и эффектами взаимодействия среди членов ошибки – SDEM. Пространственные эконометрические модели также могут быть использованы для объяснения поведения экономических агентов, отличных от географических единиц (городов, регионов, стран и т. п.), таких как физические лица, предприятия или правительства, если они связаны друг с другом через сети.

Различают три поколения пространственных эконометрических моделей в зависимости от типа данных и методов оценивания. Первое поколение состоит из статических моделей, основанных на кросс-секциях.

Второе поколение включает в себя статические модели, основанные на пространственных панельных данных. Это могут быть модели сквозной регрессии объединенных временных рядов для кросс-секций географических единиц наблюдения, но также и модели, контролирующие фиксированные и/или случайные пространственные и/или временные специфические эффекты.

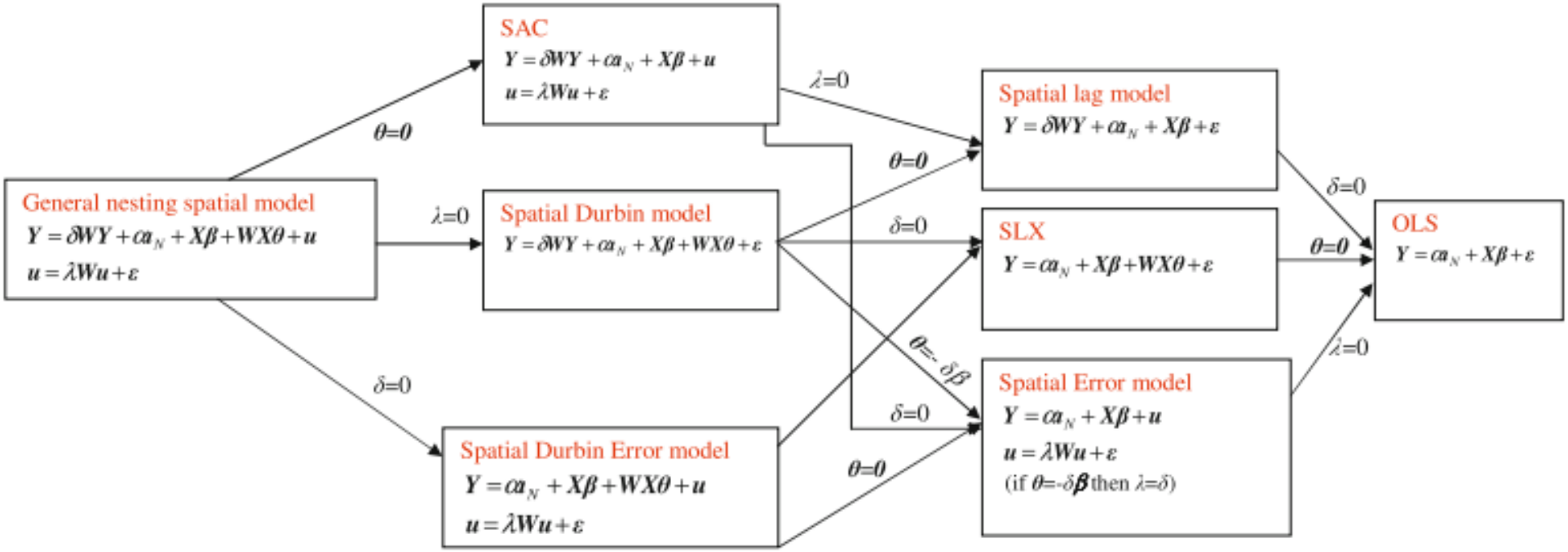

Рис. 7.1. Взаимосвязь между моделями пространственной регрессии на кросс-секциях[67]

[67] *Elhorst J.P.* Spatial Econometrics: from Cross-Sectional Data to Spatial Panels. 2014. P. 9.

Третье поколение пространственных эконометрических моделей включает динамические пространственные модели панельных данных, учитывающие пространственно-временную зависимость. Например, эффект эндогенного взаимодействия с соседней территорией в предыдущий период: $Y_{i,t} = \alpha + \gamma W_N Y_{i,t-1} + X_{i,t}\beta + \varepsilon_{i,t}$. Методы, разработанные для динамических, но непространственных моделей, и для пространственных, но нединамических моделей панельных данных, давали смещенные оценки при объединении этих методов/моделей. Поэтому сформировался пул исследований, авторы которых пытались устранить этот недостаток[68].

7.2. Модель пространственного лага

Модель пространственного лага – это метод пространственной регрессии для учета пространственной автокорреляции зависимой переменной путем включения новой переменной (в правой части уравнения), которую называют зависимой переменной с пространственным лагом. Модель пространственного лага также называется пространственной авторегрессионной моделью (Spatial Autoregressive model, SAR) или комбинированной пространственной авторегрессионной моделью регрессии. Спецификация модели соответствует общепринятым обозначениям [69]:

$$Y = \alpha + \rho \mathrm{WY} + X\beta + \varepsilon, \qquad (7.8)$$

где Y – вектор $n \times 1$ наблюдений с зависимой переменной Y, n – количество наблюдений; W – матрица $n \times n$ пространственных весов (также называется оператором пространственного лага); WY – произведение, которое также называют членом пространственного лага, представляющее дополнительную переменную (еще называется зависимой пере-

[68] *Elhorst J.P.* Spatial Econometrics: from Cross-Sectional Data to Spatial Panels. 2014. P. 95–117.

[69] *Демидова О.А.* Методы пространственной эконометрики и оценка эффективности государственных программ // Прикладная эконометрика. 2021. Т. 64. С. 107–134.

менной с пространственным лагом); ρ – коэффициент перед пространственным лагом для зависимой переменной; X – матрица $n \times k$ независимых переменных, k – количество независимых переменных; β – вектор $k \times 1$ коэффициентов; ε – вектор $n \times 1$ случайных ошибок модели.

Идея модели пространственного лага заключается в том, что на результат Y влияет не только множество независимых переменных, но также соседние значения самой зависимой переменной. Например, стоимость дома в одном муниципалитете зависит от стоимости дома в соседних местоположениях (пространственная автокорреляция), наряду с другими переменными (размер дома, количество этажей). Поэтому мы тоже должны включить это влияние в уравнение, задав соответствующую матрицу пространственных весов.

Пространственные веса – это числа, отражающие расстояние некоторого рода между целевым пространственным объектом и любым другим в пределах указанной окрестности. Если используется стандартизированная матрица пространственных весов, где сумма весов строках равна 1 (то есть $\sum_{j=1}^{n} w_{i,j} = 1$), то переменная с пространственным лагом получает средневзвешенное значение зависимых значений в соседних наблюдениях, равное I. Предположим, у нас есть пространственные объекты – муниципальные округа, изображенные на рис. 7.2[70].

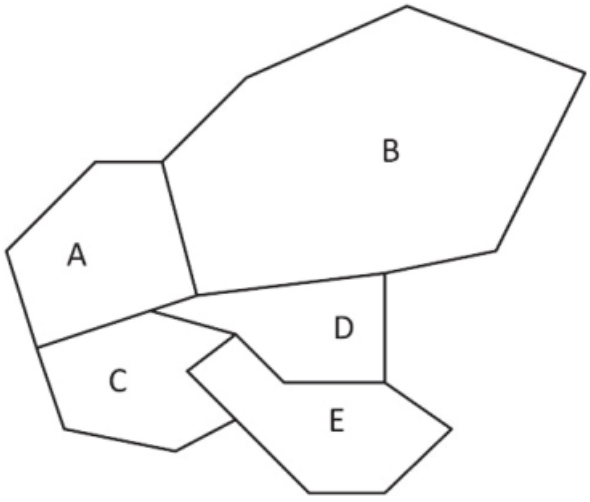

Рис. 7.2. Пример размещения полигонов для матрицы смежности

[70] *Грекусис Дж.* Методы и практика пространственного анализа. М., 2021. С. 481–482.

Матрица весов *W*, основанная на смежности согласно правилу ладьи (смежными считаются только объекты, имеющие общие границы) с учетом смежных многоугольников в качестве соседей, имеет вид:

$$Adj = \begin{bmatrix} & A & B & C & D & E & F & SUM \\ A & * & 1 & 1 & 1 & 0 & 0 & 3 \\ B & 1 & * & 0 & 1 & 0 & 0 & 2 \\ C & 1 & 0 & * & 1 & 1 & 0 & 3 \\ D & 1 & 1 & 1 & * & 1 & 1 & 5 \\ E & 0 & 0 & 1 & 1 & * & 0 & 2 \\ F & 0 & 0 & 0 & 1 & 0 & * & 1 \\ SUM & 3 & 2 & 3 & 5 & 2 & 1 & * \end{bmatrix} \quad (7.9)$$

Переменная с пространственным лагом – это взвешенная сумма соседних значений зависимой переменной *Y* для каждого местоположения в его окрестности. Для наблюдения i значение переменной Y_i с пространственным лагом определяется как:

$$[WY]_i = \sum_{j=1}^{n} w_{i,j}Y_j = w_{i,1}Y_1 + w_{i,2}Y_2 + \cdots + w_{i,n}Y_n, \quad (7.10)$$

где $w_{i,j}$ – вес между i-м и j-м объектами в матрице пространственных весов *W*.

Например, согласно рисунку 7.2, в муниципальном округе *E* значение зависимой переменной с пространственным лагом равно:

$$[WY]_E = w_{E,C}Y_C + w_{E,D}Y_D, \quad (7.11)$$

где *C* и *D* – округа с ненулевыми весами.

Модель пространственного лага можно переписать в сокращенной форме:

$$Y - \rho WY = Xb + \varepsilon. \quad (7.12)$$

Если коэффициент ρ известен, то b можно получить методом наименьших квадратов OLS. Уравнение можно также переписать как[71]:

$$Y = (I - \rho W)^{-1} Xb + (I - \rho W)^{-1} \varepsilon. \qquad (7.13)$$

Эта сокращенная форма показывает, что значение Y в каждом местоположении i определяется не только значениями X_i, но также значениями X соседей через пространственный множитель, выраженный как $(I - \rho W)^{-1}$, где I – единичная матрица.

Как показано в (7.13) значение Y в любом местоположении i является функцией от значения X в этом и в соседних местоположениях. Из-за наличия пространственной зависимости и эндогенных переменных оценки OLS нельзя использовать в пространственной модели. Поэтому следует применять альтернативные методы оценки, например, пространственный двухшаговый метод наименьших квадратов (Spatial 2-Stage Least Squares, S2SLS), метод максимального правдоподобия (ML), обобщенный метод моментов.

Как следует из названия, двухшаговый метод наименьших квадратов – это метод оценки параметров эконометрических моделей, состоящий из двух этапов (шагов), на каждом из которых применяется метод наименьших квадратов. На первом шаге каждая эндогенная переменная подвергается регрессии со всеми экзогенными переменными и инструментами (матрица инструментов); на втором шаге прогнозируемые значения, вычисленные с помощью регрессионной модели, полученной на предыдущем шаге, используются в качестве независимых переменных, заменяя эндогенные переменные. ***Эндогенные переменные*** – это независимые переменные в модели линейной регрессии, которые коррелируют с ошибкой. Наличие эндогенных переменных в регрессионной модели нарушает предположение о независимости линейной регрессии (член ошибки должен быть независимым от переменных-предикторов). ***Инструментальная переменная*** (или просто инструмент) – это переменная, сильно корреля-

[71] *Anselin L.* Spatial Econometrics. 2005. P. 18.

рующая с эндогенной переменной, но не с ошибкой. Основная идея или допущение при работе с эндогенной переменной – заменить эндогенную переменную другой (инструментальной) переменной, которая сильно коррелирует с эндогенной переменной и не коррелирует с ошибкой. Иначе говоря, если мы найдем или создадим переменную, которая сильно коррелирует с эндогенной переменной, а не с членом ошибки, то сможем применить ***двухшаговый метод наименьших квадратов (2SLS)***, чтобы заменить оригинальную переменную другой переменной. В модели пространственного лага переменная с пространственным лагом также является эндогенной. То есть модель можно выразить как:

$$Y = \rho WY + Xb + \hat{Y}\gamma + \varepsilon, \qquad (7.14)$$

где $\hat{Y}$ матрица $n \times s$ наблюдений эндогенных переменных (за исключением зависимой переменной с пространственным лагом); γ – вектор $s \times 1$ коэффициентов при эндогенных переменных.

Есть два пути выбора инструментальных переменных для эндогенных независимых переменных: это могут быть те переменные, которые успешно использовались в аналогичных исследованиях; другое решение – проверить, устраняет ли инструментальная переменная корреляцию с ошибкой.

Метод максимального правдоподобия как параметрический статистический подход можно использовать для оценки параметров моделей пространственного лага и пространственной ошибки. Статистическим предположением максимального правдоподобия является нормальное распределение членов ошибки, а пространственные веса, используемые при оценке максимального правдоподобия модели пространственного лага, должны соотноситься с симметричным отношением смежности. По этой причине оценка максимального правдоподобия модели может использоваться только в сочетании со смежностью по правилу ладьи, смежностью по правилу ферзя и диапазоном расстояний, но не с ближайшими соседями.

Результаты метода максимального правдоподобия могут быть лучше результатов метода S2SLS. Это не обязательно указывает на более точную модель. Оценка максимального правдоподобия не принимает во внимание потенциальную ненормальность распределения ошибок или их гетероскедастичность. Оценки S2SLS не требуют соблюдения предположения о нормальном распределении ошибок, лучше справляется с проблемой ненормальности распределения ошибок или их гетероскедастичности.

Оценка R-квадрат не может применяться в моделях пространственного лага. Более подходящими и сопоставимыми метриками являются логарифмическая вероятность, информационный критерий Акаике (AIC) и критерий Шварца (SW). Повышенное значение логарифма правдоподобия указывает на более высокое качество аппроксимации. И, наоборот, AIC и SW указывают на более высокое качество, когда их значения уменьшаются.

Коэффициенты имеют более сложную интерпретацию, чем в непространственной регрессии, где изменение X на одну единицу вызывает изменение Y на b единиц. В регрессии с пространственным лагом изменение X на одну единицу дает лишь часть изменения Y. Соседние значения X вызывают дополнительное изменение Y.

Назначение модели пространственного лага. Модель пространственного лага уместна, когда основное внимание уделяется пространственным взаимодействиям зависимой переменной, которая представляет собой средневзвешенное значение от значений ее соседей. Модель пространственного лага учитывает пространственную зависимость при оценке влияния и значимости независимых переменных.

7.3. Модель пространственной ошибки

Модель пространственной ошибки – это форма модели регрессии, которая учитывает пространственную зависимость включением члена пространственной ошибки. Мы включаем пространственно коррелированные ошибки из-за ненаблюдаемых особенностей или

пропущенных переменных, связанных с местоположением, когда скрытые влияния медленно изменяются по мере перемещения по территориям. Например, на решения фермеров о внедрении технологий могут влиять их соседи. Скрытые ненаблюдаемые влияния, связанные с культурой, инфраструктурой, удобствами для отдыха и множеством других факторов, для которых у нас нет доступных выборочных данных, можно объяснить, полагаясь на соседние значения, принимаемые зависимой переменной. Спецификация модели соответствует общепринятым обозначениям[72]:

$$Y = \alpha + X\beta + u,$$
$$u = \lambda W u + \varepsilon, \qquad (7.15)$$

где Y – вектор $n \times 1$ наблюдений зависимой переменной Y, n – количество наблюдений; X – матрица $n \times k$ независимых переменных, k – количество независимых переменных; β – вектор $k \times 1$ коэффициентов; W – матрица $n \times n$ пространственных весов; λ – коэффициент перед пространственным лагом для ошибки регрессии; ε – вектор $n \times 1$ специфических ошибок; эти ошибки некоррелированы, но могут проявлять гетероскедастичность.

В отличие от модели с пространственным лагом, где зависимая переменная интерпретируется как авторегрессионная, в этой модели пространственно авторегрессионной считается ошибка.

Для оценки модели пространственной ошибки обычно используются два метода: обобщенный метод моментов (Generalized Method of Moments, GMM) и метод максимального правдоподобия (Maximum Likelihood, ML).

Обобщенный метод моментов – это метод, применяемый в математической статистике и эконометрике для оценки неизвестных параметров распределений и эконометрических моделей, являющийся обобщением классического метода моментов. В отличие от классиче-

[72] *Демидова О.А.* Методы пространственной эконометрики и оценка эффективности государственных программ // Прикладная эконометрика. 2021. Т. 64. С. 107–134.

ского метода моментов количество ограничений может быть больше количества оцениваемых параметров. Обобщенный метод моментов используется для оценки коэффициента авторегрессии λ. Этот метод устойчив к гетероскедастичности и не требует выполнения предположения о нормальности распределения. Так же как в модели пространственного лага, в качестве мер соответствия используются логарифм правдоподобия, информационный критерий Акаике (AIC) и критерий Шварца (SW). Чем выше значение логарифма правдоподобия, тем лучше модель аппроксимирует фактические данные. С другой стороны, лучшей аппроксимации соответствуют меньшие значения AIC и SW. Для проверки значимости параметра авторегрессии λ используется критерий отношения правдоподобия. Если p-значение меньше уровня значимости, то нулевая гипотеза о том, что $\lambda = 0$, отклоняется. В таких случаях результат проверки считается статистически значимым, и принимается модель пространственной ошибки. Для проверки гетероскедастичности используется критерий Бройша-Пагана.

Сравнивая GMM с ML, можно утверждать, что GMM более надежен, потому что его оценки остаются действительными при наличии гетероскедастичности членов ошибки. Статистически значимые результаты, которые дает ML, могут сопровождаться гетероскедастичными остатками.

Назначение модели пространственной ошибки заключается в учете пространственной автокорреляции в остатках, соответственно, она контролирует зависимые и независимые переменные. Модель пространственной ошибки подходит, когда мы заинтересованы в коррекции пространственной автокорреляции из-за использования пространственных данных (независимо от того, является ли интересующая модель пространственной или нет). Данная модель позволяет скорректировать смещенность, вызванную пространственной автокорреляцией в пространственных данных (независимо от того, является интересующая модель пространственной или нет). По этой при-

чине иногда она выглядит более предпочтительной, потому что более надежна.

7.4. Прямые и косвенные эффекты

Применительно к модели с пространственным лагом изменение на одну единицу переменной X_k (то есть изменение на одну единицу всех значений X в векторе-столбце k для всех местоположений) приводит к общему изменению Y на $\frac{b_k}{1-\rho}$. По сути, общий эффект изменения переменной X_k на одну единицу складывается из прямого влияния b_k переменной X_k в точке i и косвенного эффекта $\frac{b_k\rho}{1-\rho}$ (пространственного множителя), обусловленного значениями переменной X_k в соседних окрестностях: $\frac{b_k}{1-\rho} = b_k + \frac{b_k\rho}{1-\rho}$, где b_k – коэффициент при переменной X_k, вычисленный с использованием модели пространственного лага.

Если обратиться к формулам (7.12) и (7.13) и перенести в левую часть уравнений пространственный лаг, то предельные эффекты изменения зависимой переменной при увеличении объясняющего фактора на единицу его измерения – это произведение коэффициента b с обратной матрицей $(I - \rho W)^{-1}$Тогда ожидаемое изменение уровня зависимой переменной $E(Y_i)$ при единичном увеличении некоторого фактора X будет равно[73]:

$$\left(\frac{\partial E(Y)}{\partial x_1} \dots \frac{\partial E(Y)}{\partial x_N}\right) = \begin{pmatrix} \frac{\partial E(Y_1)}{\partial x_1} & \dots & \frac{\partial E(Y_1)}{\partial x_N} \\ \dots & \dots & \dots \\ \frac{\partial E(Y_N)}{\partial x_1} & \dots & \frac{\partial E(Y_N)}{\partial x_N} \end{pmatrix} = (I_N - \rho W)^{-1} b. \qquad (7.16)$$

Получается, что в пространственно-эконометрических моделях изменение объясняющей переменной в каком-то регионе вызывает

[73] *Вакуленко Е.С., Ратникова Т.А., Фурманов К.К.* Эконометрика (продвинутый курс). Применение пакета Stata. М., 2020. С. 225.

изменение зависимой переменной не только в данном регионе, но и в соседних регионах, которые также влияют на данный регион. Таким образом, возникают обратные связи или что-то вроде эффекта мультипликатора. Суммируя элементы матрицы $(I - \rho W)^{-1}b$, можно рассчитать прямые, косвенные и общие эффекты. ***Прямой эффект*** определяется как среднее (по всем территориям) изменение зависимой переменной на территории при изменении объясняющей переменной в том же регионе. Другими словами, это среднее значение диагональных элементов матрицы $(I - \rho W)^{-1}b$. ***Косвенный эффект*** – это среднее изменение зависимой переменной в регионе при изменении объясняющей переменной во всех других регионах, то есть среднее значение суммы внедиагональных элементов матрицы предельных эффектов. Косвенные эффекты также называют эффектами перетока. ***Общий эффект*** – это сумма прямого и косвенного эффектов, то есть среднее изменение зависимой переменной в данном регионе в случае изменения объясняющей переменной во всех регионах.

Модель пространственной ошибки не требует расчета прямых и косвенных эффектов. Коэффициенты модели могут быть проинтерпретированы в явном виде.

7.5. Выбор спецификации пространственно-эконометрической модели

Описание основных пространственно-эконометрических моделей для кросс-секционных данных и подходов к выбору подходящего типа модели выполнено Демидовой О.А.[74] (см. табл. 7.1).

[74] *Демидова О.А.* Методы пространственной эконометрики и оценка эффективности государственных программ // Прикладная эконометрика. 2021. № 64. С. 107–134.

Таблица 7.1

Основные пространственно-эконометрические модели
для кросс-секционных данных

Тип модели	Спецификация
Модель пространственной авторегрессии (SAR)	$Y_i = \rho(WY)_i + \alpha + x_i\beta + \varepsilon_i$
Пространственная модель Дарбина (SDM)	$Y_i = \rho(WY)_i + \alpha + x_i\beta + (WX)_i\theta + \varepsilon_i$
Модель с пространственнной зависимостью в ошибках (SEM)	$Y_i = \alpha + x_i\beta + u_i,$ $u_i = \lambda(Wu)_i + \varepsilon_i$
Общая вложенная пространственная модель (GNS)	$Y_i = \rho(WY)_i + \alpha + x_i\beta + (WX)_i\theta + u_i,$ $u_i = \lambda(Wu)_i + \varepsilon_i$

Подход к выбору спецификации пространственно-эконометрической модели, предложенный Л. Анселином[75], начинается с оценивания непространственной модели линейной регрессии (по типу OLS, см. рис. 7.1). Затем выполняется проверка необходимости расширения такой модели эффектами пространственного взаимодействия. Для максимальной эффективности пространственных критериев исходная модель OLS должна обеспечивать приемлемую аппроксимацию данных. Поэтому очень важно обладать глубокими знаниями теории и практики линейной регрессии. Чтобы решить, какая модель (пространственного лага или пространственной ошибки) наиболее подходит для конкретного случая, используют пространственные критерии – множители Лагранжа, индекс Морана, критерий Анселина и Келеджиана.

[75] *Anselin L., Bera A.K., Florax R., Yoon M.J.* Simple diagnostic tests for spatial dependence // Regional Science and Urban Economics. 1996. Vol. 26(1). P. 77–104.

Коллекция ***множителей Лагранжа*** включает четыре критерия[76]:

– множитель Лагранжа, определяющий существование автокорреляции пространственного лага и необходимость включения в регрессию переменной с лагом. Если p-значение меньше уровня значимости, то в модель следует добавить переменную с лагом;

– множитель Лагранжа, определяющий существование автокорреляции пространственной ошибки и необходимость применения модели пространственной ошибки вместо OLS. Если p-значение меньше уровня значимости, то следует использовать модель пространственной ошибки;

– робастный множитель Лагранжа, определяющий существование автокорреляции пространственного лага. Если p-значение меньше уровня значимости, то это указывает на необходимость применения модели пространственного лага;

– робастный множитель Лагранжа, определяющий существование автокорреляции пространственной ошибки. Если p-значение меньше уровня значимости, то необходимо применять модель пространственной ошибки.

Робастные критерии множителя Лагранжа используются, только если оба множителя Лагранжа, лаг и ошибка, являются статистически значимыми.

Множитель Лагранжа (SARMA) указывает на использование пространственно-эконометрической модели по типу SAC, в которой присутствуют как пространственный лаг, так и пространственная ошибка. Этот критерий обычно является статистически значимым, когда изначально статистически значима модель лага или модель ошибки, поэтому он не особенно информативен и мало полезен на практике. Модель с одновременным присутствием пространственного лага и пространственной ошибки (модель по типу SAC) рекомендует-

[76] *Грекусис Дж.* Методы и практика пространственного анализа. М., 2021. С. 474–475.

ся рассматривать в том случае, когда модель с одним пространственным параметром (лагом или ошибкой) была реализована впервые и еще не была удалена пространственная автокорреляция.

Индекс Морана обнаруживает пространственную автокорреляцию, но не указывает – в каком направлении следует искать автокорреляцию – среди лагов или ошибок. Поэтому индекс Морана можно рассматривать как критерий неправильности спецификации непространственной регрессионной модели, но при этом невозможно сделать выбор в пользу модели пространственного лага или пространственной ошибки.

Критерий Анселина и Келеджиана проверяет наличие пространственной автокорреляции ошибок после применения модели пространственного лага. Вычисляется только для модели пространственного лага и не используется в модели OLS. Если p-значение меньше уровня значимости, это означает, что в модели пространственного лага сохраняется пространственная автокорреляция ошибок.

Обобщение перечисленных критериев выполнено в табл. 7.2.

На рисунке 7.3. представлен подход Л. Анселина к выбору типа пространственно-эконометрической модели (SAR или SEM) в соответствии с критериями множителя Лагранжа.

Рисунок 7.3 демонстрирует, что:

– если один из критериев множителя Лагранжа (лаг или ошибка) является статистически значимым, то выбирается соответствующая модель (лага или ошибки);

– если ни один из критериев не является значимым, оставляем результаты непространственной линейной регрессии методом наименьших квадратов (OLS);

Таблица 7.2

Статистические критерии, используемые в линейной модели регрессии для выявления потенциальных пространственных эффектов[77]

Критерий	Определяет	Проверяемая гипотеза (когда *p*-значение меньше уровня значимости, нулевая гипотеза отвергается)
Множитель Лагранжа (лаг)	Имеет место пространственная автокорреляция лага	**Нулевая гипотеза:** параметр ρ равен нулю H_0: $\rho = 0$ **Альтернативная гипотеза:** параметр ρ не равен нулю H_1: $\rho \neq 0$
Множитель Лагранжа (ошибка)	Имеет место пространственная автокорреляция ошибки	**Нулевая гипотеза:** параметр λ равен нулю H_0: $\lambda = 0$ **Альтернативная гипотеза:** параметр λ не равен нулю H_1: $\lambda \neq 0$
Робастный множитель Лагранжа (лаг)	Имеет место пространственная автокорреляция лага	**Нулевая гипотеза:** параметр ρ равен нулю H_0: $\rho = 0$ **Альтернативная гипотеза:** параметр ρ не равен нулю H_1: $\rho \neq 0$
Робастный множитель Лагранжа (ошибка)	Имеет место пространственная автокорреляция ошибки	**Нулевая гипотеза:** параметр λ равен нулю H_0: $\lambda = 0$ **Альтернативная гипотеза:** параметр λ не равен нулю H_1: $\lambda \neq 0$
Множитель Лагранжа (SARMA)	Имеет место пространственная автокорреляция лага и ошибки	**Нулевая гипотеза:** параметры ρ и λ равны нулю H_0: $\rho = 0, \lambda = 0$ **Альтернативная гипотеза:** параметры ρ и λ не равны нулю H_1: $\rho \neq 0, \lambda \neq 0$

[77] *Грекусис Дж.* Методы и практика пространственного анализа. М., 2021. С. 475–477.

Продолжение табл. 7.2

Критерий	Определяет	Проверяемая гипотеза (когда *p*-значение меньше уровня значимости, нулевая гипотеза отвергается)
Индекс Морана	Имеет место пространственная автокорреляция ошибки	**Нулевая гипотеза:** пространственная автокорреляция в ошибках отсутствует **Альтернативная гипотеза:** имеет место пространственная автокорреляция в ошибках
Критерий Анселина и Келеджиана	Имеет место пространственная автокорреляция лага (применяется только к модели пространственного лага)	**Нулевая гипотеза:** пространственная автокорреляция в ошибках отсутствует **Альтернативная гипотеза:** имеет место пространственная автокорреляция в ошибках

– если оба критерия множителя Лагранжа (лаг или ошибка) являются статистически значимыми, то следует вычислить соответствующие робастные критерии.

После вычисления робастных критериев:

– если статистически значим только один робастный критерий множителя Лагранжа (лага или ошибки), то выбираем соответствующую модель;

– если статистически значимы оба робастных критерия, то выбираем модель с большим значением робастной статистики. Иногда имеет смысл пересмотреть метод пространственной концептуализации и соответствующую матрицу пространственных весов.

Подход П. Элхорста[78] к выбору спецификации пространственно-эконометрической модели заключается в том, чтобы

[78] *Elhorst J.P.* Spatial Econometrics: from Cross-Sectional Data to Spatial Panels. 2014.

начать с более общей модели (по типу GNS, см. рис. 7.1), содержащей вложенные внутри более простые типы моделей, которые в идеале должны представлять альтернативные экономические гипотезы, требующие рассмотрения. Определяем статистическую значимость ρ – коэффициента регрессии перед пространственным лагом для зависимой переменной, λ – коэффициента регрессии перед пространственным лагом для ошибки, коэффициентов регрессии при пространственных лагах экзогенных регрессоров матрицы *X*. Если не значим только коэффициент λ, то переходим к модели по типу SDM. Если не значим коэффициент λ и параметры при пространственных лагах экзогенных регрессоров матрицы *X*, то переходим к модели по типу SAR. Если не значим только коэффициент ρ, то переходим к модели по типу SDEM. Если не значимы параметры при пространственных лагах экзогенных регрессоров матрицы *X*, то переходим к модели по типу SAC. Если не значим коэффициент ρ и параметры при пространственных лагах экзогенных регрессоров матрицы *X*, то переходим к модели по типу SEM. Если не значимы коэффициент ρ и коэффициент λ, то используем модель по типу SLX.

Отметим, что модель по типу SDM содержит пространственные лаги зависимой переменной и экзогенных регрессоров матрицы *X*. Пространственные лаги зависимой переменной (*WY*) могут аккумулировать пространственные лаги экзогенных регрессоров матрицы *X* (*WX*), не рекомендовано включать их в модель одновременно. Поэтому модель по типу SDM требует подбора инструментальных переменных для *WY* при использовании обобщенного метода моментов, и она менее популярна, чем модель по типу SAR. По этой же причине эндогенность пространственных лагов зависимой переменной, входящих в правую часть моделей SAR, SDM, GNS, SAC, требует использовать для оценивая моделей метод максимального правдоподобия и обобщенный метод моментов. В качестве инструментов для *WY* по предложению Х. Келеджиана и И. Прухи[79] используют переменные матрицы *X*, их пространственные лаги WX, W^2X, W^3X и т. д.

[79] *Kelejian H.H., Prucha I.R.* Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances // Journal of Economet-

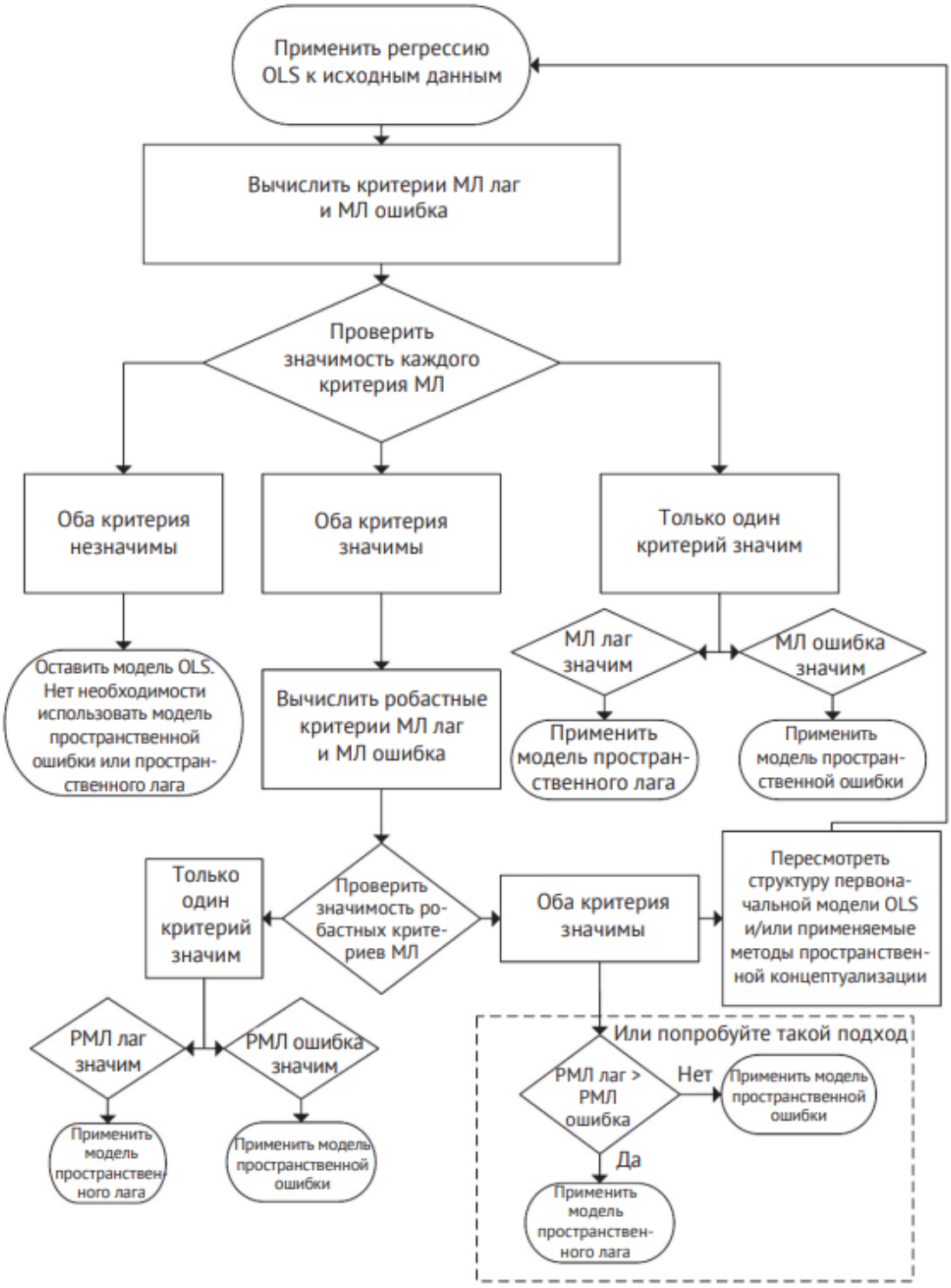

Рис. 7.3. Выбор модели пространственного лага или пространственной ошибки в соответствии с критериями множителя Лагранжа[80]

---

rics. 2010. Vol. 157(1). P. 53–67. *Kelejian H.H., Prucha I.R.* A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances // The Journal of Real Estate Finance and Economics. 1998. Vol. 17(1). P. 99–121.

[80] *Грекусис Дж.* Методы и практика пространственного анализа. М., 2021. С. 475–477.

Глоссарий

Двухшаговый метод наименьших квадратов (S2SLS) – это метод оценки параметров эконометрических моделей, состоящий из двух этапов (шагов), на каждом из которых применяется метод наименьших квадратов.

Инструментальная переменная – это переменная, сильно коррелирующая с эндогенной переменной, но не с ошибкой.

Коллекция множителей Лагранжа – статистические критерии для выбора типа пространственно-эконометрической модели.

Косвенный эффект (эффект перетока) – это среднее изменение зависимой переменной на данной территории при изменении объясняющей переменной на всех других территориях.

Критерий Анселина и Келеджиана – статистический критерий для проверки наличия пространственной автокорреляции ошибок после применения модели пространственного лага.

Метод максимального правдоподобия (ML) – параметрический статистический подход для оценки параметров моделей пространственного лага и пространственной ошибки.

Модель пространственного лага – это метод пространственной регрессии для учета пространственной автокорреляции зависимой переменной, предполагающий включение новой переменной (в правой части уравнения), которую называют зависимой переменной с пространственным лагом.

Модель пространственной ошибки учитывает пространственную автокорреляцию в остатках, соответственно, она контролирует зависимые и независимые переменные.

Обобщенный метод моментов (Generalized Method of Moments, GMM) – это метод, применяемый в математической статистике и эконометрике для оценки неизвестных параметров распределений и эконометрических моделей.

Общий эффект – это сумма прямого и косвенного эффектов, то есть среднее изменение зависимой переменной на данной территории

в случае изменения объясняющей переменной на всех других территории.

Пространственная автокорреляция – это функциональная связь между происходящим в некотором месте и в его окрестностях.

Пространственная неоднородность – это смена типа взаимосвязи в наборе данных в зависимости от местоположения.

Пространственная эконометрика – это множество методов оценки и учета пространственной зависимости (пространственной автокорреляции) и пространственной неоднородности (пространственной структуры) в регрессионном моделировании. Такие регрессионные модели также называются моделями пространственной регрессии.

Пространственные веса – это числа, отражающие расстояние некоторого рода между целевым пространственным объектом и любым другим в пределах указанной окрестности.

Прямой эффект – это среднее изменение зависимой переменной на данной территории при изменении объясняющей переменной на той же территории.

Эндогенные переменные – это независимые переменные в модели линейной регрессии, которые коррелируют с ошибкой.

Вопросы для самоконтроля

1. Что такое пространственная эконометрика?
2. Перечислите наиболее широко известные модели пространственной эконометрики?
3. Что представляет собой модель пространственного лага?
4. Что представляет собой модель пространственной ошибки?
5. Какие проблемы возникают при использовании регрессии наименьших квадратов (OLS), когда имеет место пространственная автокорреляция?
6. Перечислите основные методы оценки параметров модели пространственного лага.

7. Какие проблемы возникают при выполнении непространственного регрессионного анализа, когда имеет место пространственная неоднородность?

8. Назовите подходы к выбору типа пространственно-эконометрической модели.

9. Опишите последовательность применения критериев множителя Лагранжа для выбора типа пространственного взаимодействия.

10. Перечислите основные методы оценки параметров модели пространственной ошибки.

11. Что измеряют прямой и косвенный эффекты?

12. Что измеряет общий эффект?

ТЕМА 8. РЕГРЕССИОННЫЙ АНАЛИЗ ПАНЕЛЬНЫХ ДАННЫХ

В результате изучения данной темы обучающийся будет:

– ***знать*** определение и преимущества панельных данных; спецификации моделей панельных данных; запись спецификаций моделей панельных данных в матричном виде; спецификацию пространственно-эконометрической модели по типу SAR для панельных данных;

– ***уметь*** выбрать тип спецификации модели панельных данных и находить ее оценки; интерпретировать прямые, косвенные и общие эффекты;

– ***владеть*** навыками регрессионного анализа панельных данных и оценки пространственно-эконометрических моделей на панельных данных.

Основные вопросы:

8.1. Типы моделей регрессии на панельных данных.

8.2. Выбор типа модели.

8.3. Пространственно-эконометрическая модель SAR на панельных данных.

8.4. Примеры использования пространственно-эконометрических моделей на панельных данных.

Ключевые слова: панельные данные, модель на панельных данных, объединенная модель, модель с фиксированными эффектами, модель со случайными эффектами, тест Хаусмана

8.1. Типы моделей регрессии на панельных данных

Панельные данные состоят из наблюдений для n различных объектов в течение T различных последовательных временных периодов. Модели регрессии на панельных данных помогают

избежать смещения МНК-оценок регрессионных коэффициентов из-за наличия пропущенных переменных. Регрессионный анализ панельных данных позволяет учитывать (контролировать) некоторые виды пропущенных переменных, при которых не требуется наличия явных наблюдений. Поскольку панельные данные имеют временное и пространственное измерения, их можно записать в виде матрицы.

$$Y = \begin{bmatrix} y_{11} \\ \vdots \\ y_{1T} \\ \vdots \\ \vdots \\ y_{n1} \\ \vdots \\ y_{nT} \end{bmatrix}, X = \begin{bmatrix} x_{1,11} & \cdots & x_{k,11} \\ \vdots & \ddots & \vdots \\ x_{1,1T} & \cdots & x_{k,1T} \\ \vdots & \ddots & \vdots \\ x_{1,n1} & \cdots & x_{k,n1} \\ \vdots & \ddots & \vdots \\ x_{1,nT} & \cdots & x_{k,nT} \end{bmatrix}$$

Выделяют следующие ***преимущества использования панельных данных***[81]:

– панельные данные позволяют учитывать индивидуальную неоднородность;

– панельные данные обеспечивают меньшую коллинеарность и большую эффективность оценок;

– панельные данные предоставляют возможность изучать динамику изменений индивидуальных характеристик единиц совокупности;

– панельные данные лучше способны идентифицировать и измерить эффекты, которые не определяемы только во временных рядах или только в пространственных данных;

– панельные данные позволяют конструировать и тестировать более сложные поведенческие модели;

[81] *Ратникова Т.А.* Введение в эконометрический анализ панельных данных // Эконометрический журнал ВШЭ. 2006. № 2. С. 271–272.

– панельные данные позволяют избежать смещения, связанного с агрегированием данных;

– панельные тесты на единичный корень имеют стандартные асимптотические распределения в отличие от проблемы нестандартных распределений.

К однонаправленным моделям панельных данных относят:

– объединенную модель:

$$Y_{i_t} = \alpha + X_{i_t}\beta + \varepsilon_{i_t}. \quad (8.1)$$

– модель с фиксированными эффектами:

$$Y_{i_t} = \alpha_i + X_{i_t}\beta + \varepsilon_{i_t}, \alpha_i = z_i\alpha. \quad (8.2)$$

– модель со случайными эффектами:

$$Y_{i_t} = X_{i_t}\beta + \alpha_i + \varepsilon_{i_t}. \quad (8.3)$$

Объединенная модель предполагает, что у территорий отсутствуют индивидуальные различия зависимой переменной. Эта модель является самой ограничительной из возможных, так как предписывает одинаковое поведение всем объектам выборки во все моменты времени. В матричном виде эта модель записывается так[82]:

$$\underset{(T,1)}{y_i} = \begin{bmatrix} \underset{(T,1)}{y_1} \\ \vdots \\ \underset{(T,1)}{y_2} \\ \vdots \\ \underset{(T,1)}{y_N} \end{bmatrix} = \alpha + \underset{(K,1)}{\beta} \begin{bmatrix} \underset{(T,K)}{X_1} \\ \underset{(T,K)}{X_2} \\ \vdots \\ \underset{(T,K)}{X_i} \\ \vdots \\ \underset{(T,K)}{X_n} \end{bmatrix} + \begin{bmatrix} \underset{(T,1)}{\varepsilon_1} \\ \vdots \\ \underset{(T,1)}{\varepsilon_N} \end{bmatrix}$$

[82] *Ратникова Т.А.* Введение в эконометрический анализ панельных данных // Эконометрический журнал ВШЭ. 2006. № 2. С. 271–272.

В модели с фиксированными эффектами моделируется эффект гетерогенности анализируемой зависимой переменной между территориями. Параметр местоположения α_i независимо от времени измеряет изменение зависимой переменной на *i*-ой территории под влиянием пропущенных переменных, характеризующих индивидуальные особенности исследуемых территорий, не меняющиеся во времени. Оценки ее параметров тестируют с помощью традиционных тестов Стьюдента и Фишера. В матричном виде эта модель записывается так[83]:

$$y_{i\,(T,1)} = \begin{bmatrix} y_{1\,(T,1)} \\ \vdots \\ y_{2\,(T,1)} \\ \vdots \\ y_{N\,(T,1)} \end{bmatrix} = \beta_{(K,1)} \begin{bmatrix} X_{1\,(T,K)} \\ X_{2\,(T,K)} \\ \vdots \\ X_{i\,(T,K)} \\ \vdots \\ X_{n\,(T,K)} \end{bmatrix} + \begin{bmatrix} \vec{i}_{T\,(T,1)} & 0_{(T,1)} & \cdots & 0_{(T,1)} \\ 0_{(T,1)} & \vec{i}_{T\,(T,1)} & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ 0_{(T,1)} & \cdots & \cdots & \vec{i}_{T\,(T,1)} \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ \vdots \\ a_i \\ \vdots \\ a_N \end{bmatrix} + \begin{bmatrix} \varepsilon_{1\,(T,1)} \\ \vdots \\ \varepsilon_{N\,(T,1)} \end{bmatrix}, \text{ где } \vec{i}_{T\,(T,1)} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

В модели со случайными эффектами моделируется эффект гетерогенности зависимой переменной на территориях путем введения неизменного во времени, но специфического для каждой территории слагаемого ошибки α_i, описывающего индивидуальные различия зависимой переменной в каждой территории под влиянием ненаблюдаемых переменных. Эти различия носят случайный характер, их теоретические дисперсии предполагаются одинаковыми для всех территорий и равными σ^2_α. В матричной записи уравнение модели имеет вид[84]:

---

[83] *Ратникова Т.А.* Введение в эконометрический анализ панельных данных // Эконометрический журнал ВШЭ. 2006. № 2. С. 271–272.

[84] *Ратникова Т.А.* Введение в эконометрический анализ панельных данных // Эконометрический журнал ВШЭ. 2006. № 2. С. 280.

$$y_{i\,(T,1)} = \begin{bmatrix} y_{1\,(T,1)} \\ \vdots \\ y_{2\,(T,1)} \\ \vdots \\ y_{N\,(T,1)} \end{bmatrix} = \beta_{(K,1)} \begin{bmatrix} X_{1\,(T,K)} \\ X_{2\,(T,K)} \\ \vdots \\ X_{i\,(T,K)} \\ \vdots \\ X_{n\,(T,K)} \end{bmatrix} + \begin{bmatrix} u_{1\,(T,1)} \\ \vdots \\ u_{N\,(T,1)} \end{bmatrix}, \text{ где } u_{it} = \alpha_i + \varepsilon_{it}$$

Модель с фиксированными эффектами можно использовать не только для учета эффектов, которые меняются для разных территорий, но постоянны во времени. Ее можно также использовать для учета пропущенных переменных, которые принимают одинаковые значения для различных объектов, но изменяются во времени. Модель с временными фиксированными эффектами с одним регрессором *X* может быть записана следующим образом:

$$Y_{i_t} = X_{i_t}\beta + \lambda_t + \varepsilon_{i_t}, \qquad (8.4)$$

где λ_t – константа, которая соответствует каждому временному периоду и отражает некий эффект влияния временного периода на *Y*.

Слагаемые λ_1, …, λ_t представляют собой временные фиксированные эффекты. Изменения во временных фиксированных эффектах зачастую вызваны пропущенными переменными, которые изменяются во времени, но не меняются между территориями. Например, улучшение медицинского обслуживания с течением времени проявляется на общенациональном уровне и позволяет увеличить продолжительность жизни на всех территориях.

Если часть пропущенных переменных является постоянной во времени, но меняется между территориями (например, культурные нормы), а другая часть принимает одинаковые значения для различных территорий, но изменяется с течением времени (например, общенациональные нормы обеспеченности медицинскими работниками), то оптимальным является использование модели с включением одновременно индивидуальных (по объектам

наблюдения) и временных фиксированных эффектов. ***Модель с индивидуальными и временными фиксированными эффектами*** с одним регрессором *X* можно записать в таком виде:

$$Y_{i_t} = X_{i_t}\beta + \alpha_i + \lambda_t + \varepsilon_{i_t}, \quad (8.5)$$

где α_i представляют собой индивидуальные фиксированные эффекты, λ_t – временные фиксированные эффекты.

Модель, совмещающая в себе индивидуальные и временные фиксированные эффекты, позволяет избежать возникновения смещений в оценках, вызванных пропущенными переменными, которые могут быть постоянны как во времени, так и для различных объектов наблюдения (территорий).

Спецификации модели с фиксированными эффектами являются вариациями стандартной модели множественной регрессии, коэффициенты в рамках данных моделей могут быть оценены с помощью МНК при добавлении дополнительных бинарных переменных, соответствующих территориям и временным периодам.

8.2. Выбор типа модели

Проблема выбора моделей решается путем тестирования гипотез. ***При выборе объединенной модели против модели с фиксированными эффектами*** тестируется нулевая гипотеза об отсутствии индивидуальных эффектов. Для проверки нулевой гипотезы используется тест Чоу. Определяется наблюдаемое значение F-критерия[85]:

$$F = \frac{\frac{SS_R - SS_{UR}}{N-1}}{\frac{SS_{UR}}{NT-N-K}}; F = \frac{R_1^2}{\mathrm{v}_1} \div \frac{R_0^2}{\mathrm{v}_2},$$
$$\mathrm{v}_1 = N-1\,; \mathrm{v}_2 = NT-N-K,$$
$$F > F_{\alpha,\mathrm{v}_1,\mathrm{v}_2} \rightarrow H_1 : R_1^2 > R_0^2, \quad (8.6)$$

[85] *Елисеева И.И.* Эконометрика. М., 2014. С. 407.

где SS_R – сумма квадратов остатков в объединенной (ограниченной) модели; SS_{UR} – сумма квадратов остатков в модели с фиксированными эффектами (неограниченной модели); R_1^2 – коэффициент детерминации в модели с фиксированными эффектами; R_0^2 – коэффициент детерминации в объединенной модели; v_1, v_2 – число степеней свободы, $v_1 = N - 1$, $v_2 = NT - N - K$; N – количество панелей, T – периоды времени, K – количество параметров перед независимыми переменными.

Если вычисленное значение F-критерия окажется больше критического значения, $F > F(\alpha, N-1, NT-N-K)$, для заданного уровня значимости, то можно отклонить нулевую гипотезу и принять альтернативную гипотезу о присутствии индивидуальных эффектов, то есть сделать выбор в пользу модели с фиксированными эффектами.

При выборе объединенной модели против модели со случайными эффектами используется анализ дисперсии. Определяется наблюдаемое значение статистики Фишера[86]:

$$F = \frac{\sigma_\varepsilon^2}{T\sigma_u^2 + \sigma_\varepsilon^2} \cdot \frac{T\hat{\sigma}_b^2}{\hat{\sigma}_w^2}, \qquad (8.7)$$

где $\hat{\sigma}_b^2$ – дисперсия остатков в межгрупповой модели, $\hat{\sigma}_w^2 = \sigma_u^2$ – дисперсия остатков во внутригрупповой модели.

В случае если вычисленное значение F-критерия окажется больше критического значения, $F > F(\alpha, N - k_b, N(T-1) - k_w)$, для заданного уровня значимости, то можно отклонить нулевую гипотезу и принять альтернативную гипотезу о присутствии индивидуальных эффектов, то есть сделать выбор в пользу модели со случайными эффектами.

[86] *Елисеева И.И.* Эконометрика. М., 2014. С. 409.

Для проверки нулевой гипотезы об отсутствии индивидуальных эффектов согласно тесту множителей Лагранжа Бреуша-Пагана вычисляют LM-статистику по формуле[87]:

$$LM = \frac{NT}{2(T-1)}\left[\frac{\sum_{i=1}^{N}(\sum_{t=1}^{T}\hat{\varepsilon}_{it})^2}{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\varepsilon}_{it}^2} - 1\right]^2. \quad (8.8)$$

Если LM-статистика, вычисленная на основе остатков МНК-регрессии, больше, чем 3,84 (критического значения χ^2 с одной степенью свободы на 5 %-ном уровне значимости), то можно отклонить нулевую гипотезу и сделать выбор в пользу модели со случайными эффектами.

Для проверки нулевой гипотезы об отсутствии индивидуальных эффектов согласно тесту Хонды вычисляют статистику Хонды по формуле[88]:

$$g = \sqrt{\frac{NT}{2(T-1)}}\left[\frac{\sum_{i=1}^{N}(\sum_{t=1}^{T}\hat{\varepsilon}_{it})^2}{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\varepsilon}_{it}^2} - 1\right]^2. \quad (8.9)$$

В случае если g-статистика, вычисленная на основе остатков МНК-регрессии, больше, чем 1,64, то можно отклонить нулевую гипотезу и сделать выбор в пользу модели со случайными эффектами.

При выборе модели с фиксированными эффектами против модели со случайными эффектами тестируется нулевая гипотеза об отсутствии корреляции между индивидуальными эффектами и регрессорами (наличие случайных эффектов). Для проверки нулевой гипотезы используется тест Хаусмана. Определяется наблюдаемое значение статистики Q_H:

$$Q_H = (\hat{\beta}_{внутр} - \hat{\beta}_{FGLS})'\left[\hat{V}\left(\hat{\beta}_{внутр}\right) - V(\hat{\beta}_{FGLS})\right]^{-1}\left(\hat{\beta}_{внутр} - \hat{\beta}_{FGLS}\right),$$

[87] *Елисеева И.И.* Эконометрика. М., 2014. С. 410.
[88] *Елисеева И.И.* Эконометрика. М., 2014. С. 411.

где $\hat{\beta}_{внутр}$ – внутригрупповая оценка; $\hat{\beta}_{FGLS}$ – оценка доступного обобщенного метода наименьших квадратов.

Если Q_H-статистика больше, чем критическое значение χ^2-распределения с k_w степенями свободы, где k_w – число регрессоров во внутригрупповой модели, то можно отклонить нулевую гипотезу и сделать выбор в пользу модели с фиксированными эффектами.

8.3. Пространственно-эконометрическая модель SAR на панельных данных

Прежде всего рассмотрим общие аспекты включения пространственных эффектов в объединенные модели панельных данных и модели с фиксированными эффектами. В последних свободный коэффициент специфичен для каждой панели. В объединенную модель панельных данных пространственные эффекты вводятся путем прямого расширения модели, обобщив понятие n-мерной матрицы пространственных весов поперечного сечения Wn на размерность панели nT. Как правило, предполагается, чтобы веса оставались постоянными во времени. Следовательно, размерность $nT \times nT$ упрощается до[89]:

$$W_{nT} = I_T \otimes W_n, \tag{8.10}$$

где нижние индексы относятся к размерности матрицы, а $\otimes$ означает произведение Кронекера.

Модель пространственного лага (SAR) для объединенной панельной регрессии может быть выражена как:

$$Y = \alpha + \rho(I_T \otimes W_n)Y + X\beta + \varepsilon, \tag{8.11}$$

где ρ – параметр пространственной авторегрессии (постоянный во временном измерении).

[89] *Anselin L.* Spatial Econometrics. 2005. P. 44.

Модель SAR в сокращенной форме можно переписать как:

$$Y = [I_T \otimes (I_N - \rho W_N)^{-1}]X\beta + [I_T \otimes (I_N - \rho W_N)^{-1}]\varepsilon,$$
$$Y_t = X_t\beta + \rho W_N X_t\beta + \rho^2 W_N^2 X_t\beta + \cdots + \varepsilon_t + \rho W_N \varepsilon_t + \cdots.$$

Эта сокращенная форма показывает, что значение Y в каждом местоположении i определяется не только значениями X_i, но также значениями X соседей через пространственный множитель, выраженный как $I_T \otimes (I_N - \rho W_N)^{-1}$.

Аналогично, зависимость ошибок модели с пространственным лагом приводит к $nT \times nT$ несферической дисперсионно-ковариационной матрице ошибок в форме[90]:

$$\sum NT = \sigma_u^2 [I_T \otimes (B'_N B_N)^{-1}], \qquad (8.12)$$

где $B_n = I_n - \lambda W_n, \sigma_u^2$ – общая дисперсия ошибки, а коэффициент пространственной авторегрессии λ предполагается постоянным во временном измерении. Более сложные спецификации модели могут быть введены таким же образом.

Модель пространственного лага (SAR) для неоднородных панельных данных может быть записана как модель с детерминированным эффектом:

$$Y = \rho(I_T \otimes W_N)Y + (\alpha \otimes i_T) + X\beta + \varepsilon, \qquad (8.13)$$

где α – вектор индивидуальных фиксированных эффектов.

Важно, что для «короткой» панели, когда T мало, а N достаточно велико, пространственный фиксированный эффект (spatial fixed effects) оценивается несостоятельно, при этом оценки коэффициентов β являются состоятельными, а оценка дисперсии случайной ошибки является смещенной.

[90] *Anselin L.* Spatial Econometrics. 2005. P. 45.

Оценка моделей пространственного лага и пространственной ошибки для объединенных (однородных) панелей может быть выполнена путем прямого расширения метода максимального правдоподобия и обобщенного метода моментов. Например, рассмотрим оценку максимального правдоподобия и логарифмического правдоподобия для модели пространственного лага. Его аналог для панели требует обобщения log-Jacobian члена[91]:

$$ln|I_T \otimes (I_n - \rho W_N)| = T\, ln|I_n - \rho W_n|,$$

что дает $L = T\, ln|I_n - \rho W_n| - \frac{1}{2} ln|\Sigma_{nT}| - \frac{1}{2} \varepsilon` \Sigma_{nT}^{-1} \varepsilon$, с $\varepsilon = y - \rho(I_T \otimes W_N)y - X\beta$, и Σ_{nT} как общей $nT \times nT$ дисперсионно-ковариационной матрицей ошибок. Частным случаем, представляющим особый интерес для практики, является групповая гетероскедастичность. Это позволяет для каждого периода времени иметь отдельную дисперсию ошибки. Расширение оценки максимального правдоподобия для панельных моделей пространственных ошибок выполняется таким же образом, с использованием выражения $\sum_{nT} = \sigma_u^2 \, [I_T \otimes (B'_N B_N)^{-1}]$ в качестве дисперсионно-ковариационной матрицы ошибок.

Принципы оценки на основе инструментальных переменных и метода моментов могут быть распространены на объединенную модель панельных данных, используя преимущества пространственных весов $I_T \otimes W_n$. Например, инструментами для модели пространственного лага могут быть $(I_T \otimes W_n)X$, с X в виде матрицы $nT \times (k - 1)$, исключая константу. Оценки обобщенного метода моментов могут быть обобщены на объединенную модель панельных данных путем замены пространственных весов для отдельного уравнения на их панельные аналоги, в частности, для модели по типу SAR:

$$\varepsilon = \lambda(I_T \otimes W_n)\varepsilon + u, \qquad (8.14)$$

где ε, u – векторы размерности $nT \times 1$, а $u \sim iid[0, \sigma^2_u I_{nT}]$.

[91] *Anselin L.* Spatial Econometrics. 2005. P. 45.

Подходы к тестированию спецификации объединенной модели панельных данных формируются непосредственно из моделей на кросс-секциях. К примеру, статистику LM множителей Лагранжа для выбора типа пространственного взаимодействия – в лаге или в ошибке, можно получить следующим образом[92]:

$$LM_{\lambda} = \frac{\left[\frac{e'(I_T \otimes W_n)e}{\frac{e'e}{nT}}\right]^2}{Ttr(W_nW_n + W'_nW_n)},$$

$$LM_{\rho} = \frac{\left[\frac{e'(I_T \otimes W_N)y}{\frac{e'e}{nT}}\right]^2}{\left[\frac{(W\hat{y})'M(W\hat{y})}{\hat{\sigma}^2}\right] + Ttr(W_nW_n + W'_nW_n)},$$

где e – вектор МНК-остатков размерности $nT \times 1$, $W\hat{y} = (I_T \otimes W_n)X\hat{\beta}$ – пространственный лаг предсказанных оценок регрессии, $M = I_{nT} - X(X'X)^{-1}X''$.

Обе статистики распределены асимптотически как χ^2, имеют постоянный во времени пространственный параметр. Тесты множителей Лагранжа для выбора между линейной моделью и SEM для панельных данных изложены Б. Балтаджи[93].

8.5. Примеры использования пространственно-эконометрических моделей на панельных данных

В работе «Пространственный анализ конвергенции регионов России» О.С. Балаш[94] изучает σ и β-конвергенцию регионов России

---

[92] *Anselin L.* Spatial Econometrics. 2005. P. 46.

[93] *Baltagi B.H., Song S.H., Koh W.* Testing panel data regression models with spatial error correlation // Journal of Econometrics. 2003. Vol. 117(1). P. 123–150.

[94] *Балаш О.С.* Пространственный анализ конвергенции регионов России // Известия Саратовского университета: Экономика. Управление. Право. 2012. № 12(4). С. 45–52.

по темпам роста ВРП. Исследование охватывает период с 1995 года по 2010 год и доступные показатели 78 регионов. Для тестирования σ-конвергенции был построен временной ряд коэффициента вариации ВРП на душу населения. Для обнаружения β-конвергенции были протестированы эконометрические модели, соответствующие модели Солоу, безусловной и условной β-конвергенции с пространственным лагом в зависимой переменной (SAR), с пространственным лагом в ошибке (SEM) и по типу Дарбина и по типу SAC. Статистический анализ не подтвердил σ-конвергенцию, но показал β-конвергенцию и подтвердил наличие пространственной зависимости для экономического роста регионов России. Из результатов следует, что для получения несмещенных эконометрических оценок необходимо учитывать географический фактор при исследовании социально-экономического развития страны.

В другой работе «Пространственно-авторегрессионная модель для двух групп взаимосвязанных регионов (на примере восточной и западной части России)» Демидова О.А.[95] предлагает обобщение модели пространственной авторегрессии для случая, когда рассматриваемые регионы разбиты на две группы (восточные и западные регионы), влияющие друг на друга. В исследовании в качестве зависимых переменных в оцениваемых моделях выбраны: уровень безработицы, относительная заработная плата (отношение числа минимальных потребительских корзин, которые может купить на свою заработную плату потребитель в регионе и по России в среднем), рост ВРП в регионе за год. Расчеты выполнены по данным за 2000–2010 годы для 75 регионов России. Автор использует расщепленную на четыре части взвешивающую матрицу, модификацию модели SAR и получает вывод о том, что позитивные изменения, происходящие в западных регионах, обычно положительно влияют на восточные ре-

[95] *Демидова О.А.* Пространственно-авторегрессионная модель для двух групп взаимосвязанных регионов (на примере восточной и западной части России) // Прикладная эконометрика. 2014. № 34(2). С. 19–35.

гионы, а любые изменения, происходящие в восточных регионах, не оказывают влияния на западные.

В работе «Методы пространственной эконометрики и оценка эффективности государственных программ» О.А. Демидова[96] систематизирует работы по пространственно-эконометрическому моделированию российских региональных показателей и заключает, что «пространственно-эконометрические модели были существенно модифицированы для отражения российских реалий». Специфика модификации распространяет такие модели на оценку эффективности государственных программ.

В работе «Анализ связи между региональными рынками труда в России с использованием модели Оукена» Е.С. Вакуленко[97] оценивает коэффициент Оукена на панельных данных 78 российских регионов за период 1998–2013 гг., выделяя прямые и косвенные эффекты, вводит понятия коэффициентов самостоятельности и влияния для ранжирования регионов при проведении политики на рынке труда.

В исследовании “Spatial Analysis of Regional Productivity Based on β-Convergence Models” Е.И. Кадочникова, Ю.А. Варламова, Ю.С. Колесникова[98] на панельных данных российских регионов с 2009 по 2018 годы выявили пространственную положительную корреляцию производительности труда, а между темпами роста реальных затрат на технологические инновации – пространственную отрицательную корреляцию (сильные регионы «стягивают» инновации со слабых соседей), для обнаружения предполагаемой β-конвергенции производительности

[96] *Демидова О.А.* Методы пространственной эконометрики и оценка эффективности государственных программ // Прикладная эконометрика. 2021. № 64. С. 107–134.

[97] *Вакуленко Е.С.* Анализ связи между региональными рынками труда в России с использованием модели Оукена // Прикладная эконометрика. 2015. № 40(4). С. 28–48.

[98] *Kadochnikova E, Varlamova Y, Kolesnikova J.* Spatial Analysis of Regional Productivity Based on Beta-Convergence Models // Montenegrin Journal of Economics. 2022. Vol. 18, Is. 3. P. 133–143.

в регионах применили пространственно-эконометрические модели SAR, SEM, SDM.

В работе «Влияние численности занятых на заработную плату и цены на жилье в российских регионах» Гильтман М.А., Антосик Л.В., Варламова Ю.А., Ларионова Н.И.[99] исследуют влияние численности занятых на заработную плату и цены жилья в субъектах с использованием пространственно-эконометрических панельных моделей с фиксированными эффектами. Исследование выявило пространственные зависимости между локальными и между региональными рынками труда, показало, что изменение численности занятых в регионе значимо положительно влияет на цены жилья и значимо отрицательно – на реальную заработную плату. Авторы рекомендуют использовать результаты анализа для проведения социальной, региональной и миграционной политики.

В исследовании «Конвергенция экономического роста и цифровизация домохозяйств: пространственный анализ взаимосвязи на региональных панельных данных» Кадочникова Е.И.[100] обнаруживает положительное влияние цифровой инфраструктуры домохозяйств на средний темп экономического роста в регионах с учетом пространственных зависимостей. Автор выявляет условную β-конвергенцию средних темпов роста валового регионального продукта как в краткосрочной, так и в долгосрочной перспективе, подтверждает вывод Солоу об убывающей отдаче избыточных факторов производства. В аналогичном исследовании “Savings Rates and Consumption Convergence in Regions: Spatial Analysis” Багаутдинова Н.Г. и

[99] *Гильтман М.А., Антосик Л.В., Варламова Ю.А., Ларионова Н.И.* Влияние численности занятых на заработную плату и цены на жилье в российских регионах // Вопросы экономики. 2022. № 8. С. 95–117.

[100] *Кадочникова Е.И.* Конвергенция экономического роста и цифровизация домохозяйств: пространственный анализ взаимосвязи на региональных панельных данных // Актуальные проблемы экономики и права. 2020. № 3. С. 487–507.

Кадочникова Е.И.[101] выявляют пространственную кооперацию потребления и β-конвергенцию средних темпов роста потребления в долгосрочной перспективе с помощью моделей типа SAR, SEM, SAC на панельных данных регионов России с 2014 по 2019 годы, показывают негативное влияние нормы сбережений на средний темп роста потребления на душу населения и отсутствие влияния переменных цифровизации. Авторы рекомендуют использовать результаты исследования в практической деятельности при реализации концепции устойчивого развития регионов на основе институционального подхода с учетом пространственной дифференциации.

Представленные примеры использования пространственно-эконометрических моделей на панельных данных демонстрируют востребованность и перспективность статистического анализа пространственных данных, растущий интерес к моделированию российских региональных показателей с использованием пространственно-эконометрического инструментария.

Глоссарий

Модель с фиксированными эффектами – регрессионная модель панельных данных, в которой моделируется эффект гетерогенности между объектами наблюдения с инвариантным по отношению ко времени, но специфическим для каждого объекта наблюдения параметром местоположения α_i.

Модель со случайными эффектами – регрессионная модель панельных данных, в которой моделируется эффект гетерогенности объектов наблюдения путем введения неизменного во времени, но специфического для каждого объекта наблюдения слагаемого ошибки α_i, которое предполагается случайным, независимым от оставшейся части ошибки u_{it}.

[101] *Bagautdinova N., Kadochnikova E.* Savings Rates and Consumption Convergence in Regions: Spatial Analysis // Industrial Engineering and Management Systems. 2022. Vol. 21, Is. 2. P. 228–237.

Объединенная модель – регрессионная модель панельных данных, которая предписывает одинаковое поведение всем объектам выборки во все моменты времени.

Панельные данные – множество данных, состоящих из наблюдений за однотипными статистическими объектами, в течение нескольких временных периодов.

Сбалансированная панель – это набор данных, в котором каждый объект (территория) панели наблюдается каждый временной период.

Тест множителей Лагранжа Бреуша-Пагана – эконометрический тест для выбора между объединенной моделью панельных данных и моделью со случайными эффектами. Нулевая гипотеза в пользу объединенной модели.

Тест Фишера (для панелей) – эконометрический тест для выбора между объединенной моделью панельных данных и моделью с фиксированными эффектами. Нулевая гипотеза в пользу объединенной модели.

Тест Хаусмана – эконометрический тест для выбора между моделью со случайными эффектами и моделью с фиксированными эффектами. Нулевая гипотеза в пользу модели со случайными эффектами.

Тест Хонды – эконометрический тест для выбора между объединенной моделью панельных данных и моделью со случайными эффектами. Нулевая гипотеза в пользу объединенной модели.

Тест Чоу – эконометрический тест для выбора между объединенной моделью панельных данных и моделью с фиксированными эффектами. Нулевая гипотеза в пользу объединенной модели.

Вопросы для самоконтроля

1. Назовите отличие панельных данных от кросс-секций и преимущества панельных данных.
2. Что представляет собой объединенная модель?
3. Что представляет собой модель с фиксированными эффектами?

4. Какими способами можно оценить модель с фиксированными эффектами?

5. Что представляет собой модель со случайными эффектами?

6. Чем модель со случайными эффектами отличается от модели с фиксированными эффектами?

7. Назовите тесты для выбора типа регрессионной модели на панельных данных.

8. В чем состоит нулевая гипотеза теста Хаусмана?

9. В чем состоит нулевая гипотеза теста Бреуша-Пагана?

10. Как связаны между собой LM-статистика в тесте множителей Лагранжа Бреуша-Пагана и g-статистика в тесте Хонды?

11. Запишите спецификацию пространственно-эконометрической модели по типу SAR для панельных данных.

12. Перечислите основные методы оценки параметров пространственно-эконометрических моделей на панельных данных.

13. Какая модификация модели SAR была предложена О.А. Демидовой в статье «Пространственно-авторегрессионная модель для двух групп взаимосвязанных регионов (на примере восточной и западной части России)»?

14. Какие типы пространственно-эконометрических моделей на панельных данных были применены в статье О.С. Балаш «Пространственный анализ конвергенции регионов России»?

15. Какие подходы для пространственного анализа эффекта воздействия описаны в статье О.А. Демидовой «Методы пространственной эконометрики и оценка эффективности государственных программ»?

Литература

1. *Акулич И.Л.* Математическое программирование в примерах и задачах / И.Л. Акулич. – М.: Высшая школа, 1986. – 319 с.

2. *Балаш В.А.* Пространственная корреляция в статистических исследованиях / В.А. Балаш, А.Р. Файзлиев // Вестник Саратовского государственного социально-экономического университета. – 2008. – № 4. – С. 122–125.

3. *Балаш О.С.* Пространственный анализ конвергенции регионов России / О.С. Балаш // Известия Саратовского университета: Экономика. Управление. Право. – № 12(4). – 2012. – С. 45–52.

4. *Боровиков В.* STATISTICA. Искусство анализа данных на компьютере: Для профессионалов / В. Боровиков. – СПб.: Питер, 2003. – 688 с.

5. *Вакуленко Е.С.* Анализ связи между региональными рынками труда в России с использованием модели Оукена / Е.С. Вакуленко // Прикладная эконометрика. – 2015. – №. 4 (40). – С. 28–48.

6. *Вакуленко Е.С.* Эконометрика (продвинутый курс). Применение пакета Stata: учебное пособие для вузов / Е.С. Вакуленко, Т.А. Ратникова, К.К. Фурманов. – М.: Юрайт, 2020. – 246 с.

7. *Гильтман М.А.* Влияние численности занятых на заработную плату и цены на жилье в российских регионах / М.А. Гильтман, Л.В. Антосик, Ю.А. Варламова, Н.И. Ларионова // Вопросы экономики. – 2022. – № 8. – С. 95–117.

8. *Грекусис Дж.* Методы и практика пространственного анализа / Дж. Грекусис. – М.: ДМК Пресс, 2021. – 540 с.

9. *Гурьянова Л.С.* Методы и модели анализа пространственной кластеризации темпов социально-экономического развития регионов / Л.С. Гурьянова, Г.А. Холодный, А.С. Лукьянчикова // Проблемы экономики. – 2013. – № 2. – С. 242–250.

10. *Демидова О.А.* Выявление пространственных эффектов для основных макроэкономических показателей российских регионов / О.А. Демидова // НИУ ВШЭ. – 2013. – 26 с. URL:

https://economics.hse.ru/data/2013/12/03/1335971579/Demidova_Article_HSE_2013.pdf (дата обращения: 27.01.2023).

11. *Демидова О.А.* Методы пространственной эконометрики и оценка эффективности государственных программ / О.А. Демидова // Прикладная эконометрика. – 2021. – Т. 64. – С. 107–134.

12. *Демидова О.А.* Пространственно-авторегрессионная модель для двух групп взаимосвязанных регионов (на примере восточной и западной части России) / О.А. Демидова // Прикладная эконометрика. 2014. – № 34 (2). – С.19–35.

13. *Демидова О.А.* Пространственно-эконометрическое моделирование экономического роста российских регионов: имеют ли значение институты? / О.А. Демидова, Э. Камалова // Экономическая политика. – 2021. – Т. 16. – № 2. – С. 34–59.

14. *Демидова О.А.* Эконометрика: учебник и практикум для вузов / О.А. Демидова, Д.И. Малахов. – М.: Издательство Юрайт, 2022. – 334 с.

15. *Елисеева И.И.* Общая теория статистики: учебник / Под ред. И.И. Елисеевой. – 5-е изд., перераб. и доп. – М.: Финансы и статистика, 2004. – 656 с.

16. *Елисеева И.И.* Эконометрика: учебник для магистров / И.И. Елисеева; под редакцией И.И. Елисеевой. – М.: Юрайт, 2014. – 449 с.

17. *Жаворонков А.В.* Результаты применения коэффициентов корреляции Кендалла для выявления определенных параметров / А.В. Жаворонков, А.Л. Королёв // Модернизация отечественной системы управления: анализ тенденций и прогноз развития: материалы Всероссийской научно-практической конференции и XII–XIII Дридзевских чтений. – М.: Институт социологии Российской академии наук, 2014. – С. 191–196.

18. *Кадочникова Е.И.* Конвергенция экономического роста и цифровизация домохозяйств: пространственный анализ взаимосвязи на региональных панельных данных / Е.И. Кадочникова // Актуальные проблемы экономики и права. – 2020. – № 3. – С. 487–507.

19. *Картаев Ф.* Введение в эконометрику: учебник / Ф. Картаев. – М.: Экономический факультет МГУ имени М.В. Ломоносова, 2019. – 472 с.

20. *Кремер Н.Ш.* Теория вероятностей и математическая статистика: учебник и практикум для вузов / Н.Ш. Кремер. – 3-е изд., перераб. и доп. – М.: Юнити, 2012. – 551 с.

21. *Кремер Н.Ш.* Эконометрика: учебник для студентов вузов / Н.Ш. Кремер, Б.А. Путко; под ред. Н.Ш. Кремера. – 3-е изд., перераб. и доп. – М.: ЮНИТИ-ДАНА, 2012. – 328 с.

22. *Лурье И.К.* Информатика с основами геоинформатики. Часть 2. Основы геоинформатики / И.К. Лурье, Т.Е. Самсонов. – М.: Географический факультет МГУ, 2016. – 200 с.

23. *Наумов И.В.* Методологический подход к моделированию и прогнозированию воздействия пространственной неоднородности процессов распространения COVID-19 на экономическое развитие регионов России / И.В. Наумов, Ю.С. Отмахова, С.С. Красных // Компьютерные исследования и моделирование. – 2021. – Т. 13. – № 3. – С. 629–648.

24. *Наумов И.В.* Цифровизация промышленного производства в регионах России. Пространственные взаимосвязи / И.В. Наумов, Ю.В. Дубровская, Е.В. Козоногова // Экономика региона. – 2020. – Т. 16. – Вып. 3. – С. 896–910.

25. *Окунев И.Ю.* Основы пространственного анализа: монография / И.Ю. Окунев. – М.: Издательство «Аспект Пресс», 2020. – 255 с.

26. *Орлов А.И.* Вероятностно-статистические модели корреляции и регрессии / А.И. Орлов // Научный журнал КубГАУ. – 2020. – № 160. – С. 1–3.

27. *Орлов А.И.* Вероятность и прикладная статистика – основные факты: учебное пособие / А.И. Орлов. – М.: КНОРУС, 2010. – 95 с.

28. Основы регрессионного анализа // ArcMap URL: https://desktop.arcgis.com/ru/arcmap/latest/tools/spatial-statistics-toolbox/regression-analysis-basics.htm (дата обращения: 11.11.2022).

29. *Павлов Ю.В.* Пространственные взаимодействия: оценка на основе локального и глобального индексов Морана / Ю.В. Павлов, Е.Н. Королёва // Пространственная экономика. – 2014. – № 3. – С. 95–109.

30. Приказ Федерального агентства по техническому регулированию и метрологии от 22 сентября 2016 г. N 1189-ст «О введении в действие межгосударственного стандарта» (ГОСТ 33707-2016 (ISO/IEC 2382:2015)). URL: https://base.garant.ru/71572028/ (дата обращения: 11.11.2022).

31. *Ратникова Т.А.* Введение в эконометрический анализ панельных данных / Т.А. Ратникова // Эконометрический журнал ВШЭ. – 2006. – № 2. – С. 267–316.

32. *Самсонов Т.* Пространственная регрессия. Пространственная статистика и моделирование на языке R / Т. Самсонов. URL: https://tsamsonov.github.io/r-spatstat-course/spreg.html (дата обращения: 10.12.2022).

33. Статистика: учебник для вузов / под редакцией И.И. Елисеевой. – 3-е изд., перераб. и доп. – М.: Издательство Юрайт, 2023. – 361 с.

34. *Черненко В.Д.* Высшая математика в примерах и задачах: учебное пособие для вузов / В.Д. Черненко. В 3 т.: Т. 3. – 2-е изд., перераб. и доп. – СПб.: Политехника, 2011. – 507 с.

35. Anselin L. An Introduction to Spatial Data Analysis. URL: http://geodacenter.github.io/workbook/5a_global_auto/lab5a.html#fn1 (date of access: 18.01.2023).

36. *Anselin L.* Simple diagnostic tests for spatial dependence / L. Anselin, A.K. Bera, R. Florax, M.J. Yoon // Regional Science and Urban Economics. – 1996. – Vol. 26 (1). – P. 77–104.

37. *Anselin L.* Spatial Econometrics / L. Anselin. University of Illinois, Urbana-Champaign. – 2005. – 75 p.

38. *Bagautdinova N.* Savings Rates and Consumption Convergence in Regions: Spatial Analysis / N. Bagautdinova, E. Kadochnikova // Industrial Engineering and Management Systems. – 2022. –Vol. 21, Is. 2. – P. 228–237.

39. *Dubrovskaya J.V.* The impact of digitalization on the demand for labor in the context of working specialties: Spatial analysis / J.V. Dubrovskaya, E.V. Kozonogova // Вестник СПбГУ. Экономика. – 2021. – № 37(3). – С. 395–412.

40. *Elhorst J.P.* Spatial Econometrics: from Cross-Sectional Data to Spatial Panels / J.P. Elhorst. – Springer, 2014. – 119 p.

41. *ESRI*. Матрица диаграммы рассеяния. URL: https://clck.ru/32nSN4 (дата обращения: 27.11.2022).

42. GADM. URL: https://gadm.org/ (дата обращения: 27.11.2022).

43. *Griffith D.A.* Spatial Regression Models / D.A. Griffith, Y. Chun // In: Huang, B. (ed). Comprehensive geographic information systems. Elsevier, 2018. – Vol. 3. – P. 1–27.

44. *Kadochnikova E.* Spatial Analysis of Regional Productivity Based on Beta-Convergence Models // E. Kadochnikova, Y. Varlamova, J. Kolesnikova // Montenegrin Journal of Economics. – 2022. – Vol. 18, Is. 3. – P. 133–143.

45. *Kelejian H.H.* A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances / H.H. Kelejian, I.R. Prucha // The Journal of Real Estate Finance and Economics. – 1998. – Vol. 17(1). – P. 99–121.

46. *Kelejian H.H.* Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances / H.H. Kelejian, I.R. Prucha // Journal of Econometrics. – 2010. – Vol. 157(1). – P. 53–67.

47. *Krugman P.R.* Increasing Returns and Economic Geography / P.R. Krugman // The Journal of Political Economy. – 1991. – Vol. 99. – № 3. – P. 483–499.

48. *LeSage J.P.* An Introduction to Spatial Econometrics / J.P. LeSage // Revue D Économie Industrielle. – 2008. – Vol. 123(123). – P. 19–44.

49. *Maity R.* Basic concepts of probability and statistics / R. Maity // Statistical Methods in Hydrology and Hydroclimatology. – Springer, Singapore, 2022. – P. 7-49.

50. *Plümper T.* Model specification in the analysis of spatial dependence / T. Plümper, E. Neumayer // European Journal of Political Research. – 2010. – Vol. 49(3). – P. 418–442.

51. *Stakhovych S.* Specification of spatial models: A simulation study on weights matrices / S. Stakhovych, T.H. Bijmolt // Papers in Regional Science. – 2009. – Vol. 88(2). – P. 389–408.

52. Varlamova J. Labor productivity in the digital era: a spatial-temporal analysis / J. Varlamova, N. Larionova // International Journal of Technology. – 2020. – Vol.11(6). – P. 1191–1200.

ДЛЯ ЗАПИСЕЙ

Учебное издание

Кадочникова Екатерина Ивановна
Варламова Юлия Андреевна

СТАТИСТИЧЕСКИЙ АНАЛИЗ ПРОСТРАНСТВЕННЫХ ДАННЫХ

Учебное пособие

Подписано в печать 06.04.2023.
Бумага офсетная. Печать цифровая.
Формат 60x84 1/16. Гарнитура «Times New Roman».
Усл. печ. л. 8,14. Уч.-изд. л. 4,72. Тираж 100 экз. Заказ 48/3

Отпечатано в типографии
Издательства Казанского университета

420008, г. Казань, ул. Профессора Нужина, 1/37
тел. (843) 206-52-14 (1704), 206-52-14 (1705)