

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**

Кафедра теории функций и стохастического анализа

**ПРОСТРАНСТВЕННЫЕ РЕГРЕССИОННЫЕ МОДЕЛИ С
ДИСКРЕТНОЙ СОСТАВЛЯЮЩЕЙ**

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студента 4 курса 412 группы

направления 01.03.02 — Прикладная математика и информатика

механико-математического факультета

Федорова Никиты Валерьевича

Научный руководитель

старший преподаватель

А. Д. Луньков

Заведующий кафедрой

д. ф. -м. н., доцент

С. П. Сидоров

Саратов 2025

СОДЕРЖАНИЕ

ВВЕДЕНИЕ	3
1 Основное содержание работы	6
ЗАКЛЮЧЕНИЕ	11

ВВЕДЕНИЕ

Современные социально-экономические исследования все чаще сталкиваются с необходимостью анализа данных, имеющих пространственную природу и дискретный характер. Особую актуальность такие модели приобретают при изучении региональных процессов, где наблюдаемые явления часто зависят не только от внутренних характеристик территорий, но и от их взаимного расположения и взаимодействия.

Пространственные регрессионные модели с дискретной составляющей представляют собой мощный инструмент для анализа подобных данных, позволяя учитывать как качественную природу зависимых переменных, так и пространственную автокорреляцию между наблюдениями. Развитие этих методов в последние десятилетия открыло новые возможности для моделирования миграционных процессов, анализа рынка труда, изучения распространения инноваций и других социально-экономических феноменов.

Цель работы

Целью данной работы является исследование пространственных регрессионных моделей с дискретной составляющей, включая анализ их теоретических основ, методов оценки и практического применения для изучения региональных социально-экономических процессов.

Задачи исследования

- Изучить теоретические основы парного и множественного регрессионного анализа
- Рассмотреть модели бинарного и множественного выбора, включая их спецификации и методы оценки
- Проанализировать модели с урезанными и цензурированными выборками
- Исследовать модели "времени жизни" и их применение в экономических исследованиях
- Разработать и оценить пространственные регрессионные модели для анализа региональных данных

- Провести практическое применение моделей на реальных данных российских регионов

В первом разделе представляет фундаментальные основы регрессионного анализа. В главе подробно рассматривается метод наименьших квадратов, включая вывод формул для оценок коэффициентов, анализ их статистических свойств (несмещенность, эффективность, состоятельность). Особое внимание уделяется МНК, анализу качества модели через коэффициент детерминации и стандартные ошибки, построению доверительных интервалов и процедурам проверки гипотез. Глава завершается критическим анализом ограничений парной регрессии, обосновывающим необходимость перехода к множественным моделям.

Второй раздел расширяет рассмотренные методы на случай нескольких независимых переменных. В главе представлена геометрическая интерпретация множественной регрессии, матричный вывод оценок МНК, анализ проблемы мультиколлинеарности и методов ее диагностики. Подробно обсуждаются вопросы спецификации модели, включая тесты на включение/исключение переменных, использование скорректированного коэффициента детерминации. Особое внимание уделяется интерпретации коэффициентов в множественной регрессии и сравнению с парным случаем.

Третий раздел посвящен анализу ситуаций, когда зависимая переменная является качественной. В главе последовательно рассматриваются: линейная модель вероятности с анализом ее недостатков; probit- и logit-модели с выводом функций правдоподобия; модели множественного выбора и их ограничения, связанные с предположением о независимости альтернатив; модели упорядоченного выбора для ранжированных данных. Для каждого типа моделей приводятся методы оценки параметров и критерии выбора между альтернативными спецификациями.

В четвертом разделе рассматриваются подходы к работе с неполными данными. Анализируются tobit-модель, модель Хекмана с двухшаговой процедурой оценки. Подробно обсуждаются различия между урезанными и цензурированными выборками, методы коррекции смещения отбора. Для каждой модели приводятся примеры практического применения в экономических исследованиях, обсуждаются проблемы идентификации параметров.

Пятый раздел фокусируется на анализе продолжительности до наступления события. В главе рассматриваются: параметрические модели (экспоненциальная, Вейбулла) с анализом их функций опасности; методы оценки параметров для цензурированных данных; непараметрические подходы (оценка Каплана-Мейера). Особое внимание уделяется интерпретации коэффициентов и выбору между альтернативными спецификациями на основе анализа функций выживания и опасности.

Шестой раздел представляет современные подходы к учету пространственной автокорреляции. Детально анализируются: SAR модели с лагом зависимой переменной; SEM модели с пространственной корреляцией ошибок; SDM модели с лагами независимых переменных. Для каждого типа рассматриваются методы оценки (MLE, SML, байесовские подходы), проблемы вычислительной реализации и интерпретации коэффициентов. Приводятся примеры построения матриц пространственных весов и тестирования на пространственную зависимость.

Седьмой раздел содержит эмпирическое исследование демографических процессов в российских регионах. Подробно описываются: источники данных и методика формирования переменных; результаты оценки классической и пространственной logit-моделей; анализ значимости социально-экономических факторов и силы пространственных эффектов.

1 Основное содержание работы

Парный регрессионный анализ представляет собой фундаментальный метод эконометрического моделирования, позволяющий исследовать статистическую зависимость между двумя переменными. Это выражается через линейную модель:

$$y_i = \alpha + \beta x_i + u_i, \quad i = 1, \dots, n \quad (1)$$

где y_i – зависимая переменная, x_i – регрессор, α и β – параметры модели, подлежащие оценке, а u_i – стохастическое возмущение, включающее в себя совокупное влияние неучтенных факторов.

Множественный регрессионный анализ представляет собой естественное расширение парной регрессии на случай нескольких независимых переменных. Формально модель записывается как:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i, \quad i = 1, \dots, n, \quad (2)$$

где y_i – зависимая переменная, x_{ji} – независимые переменные, β_j – параметры модели, ε_i – случайная ошибка.

Линейная модель вероятности позволяет оценить вероятность наступления бинарного события с помощью линейной регрессии. Но данная модель имеет значительный недостаток: предсказанные значения вероятностей могут выходить за пределы допустимого интервала $[0, 1]$. Для решения этой проблемы используется функция $F(\cdot)$, которая ограничивает возможные значения вероятности в пределах от 0 до 1. Таким образом:

$$P(y_t = 1) = F(x'_t \beta), \quad (3)$$

где $F(\cdot)$ – это функция распределения случайной величины.

Далее введем скрытую переменную y_t^* , которая связана с независимыми переменными x_t следующим уравнением:

$$y_t^* = x'_t \beta + \varepsilon_t, \quad (4)$$

где ε_t – случайная ошибка, нормально распределённая с нулевым сред-

ним и дисперсией σ^2 . Переменная y_t^* называется латентной, поскольку её нельзя наблюдать напрямую. Таким образом:

$$\begin{cases} y_t = 1, & \text{если } y_t^* \geq 0 \\ y_t = 0, & \text{если } y_t^* < 0 \end{cases} \quad (5)$$

В контексте анализа региональных данных ключевым элементом становится пространственная зависимость. Она проявляется в том, что значения переменной в одном регионе могут зависеть от значений в соседних.

Для моделирования таких эффектов используются:

— **SAR (Spatial Autoregressive Model):**

$$y = \rho W y + X\beta + \varepsilon, \quad (6)$$

где W — матрица пространственных весов, ρ — коэффициент пространственной автокорреляции.

— **SEM (Spatial Error Model):**

$$y = X\beta + u, \quad u = \lambda W u + \varepsilon, \quad (7)$$

где пространственная зависимость содержится в остаточном компоненте.

— **SDM (Spatial Durbin Model):**

$$y = \rho W y + X\beta + W X\theta + \varepsilon, \quad (8)$$

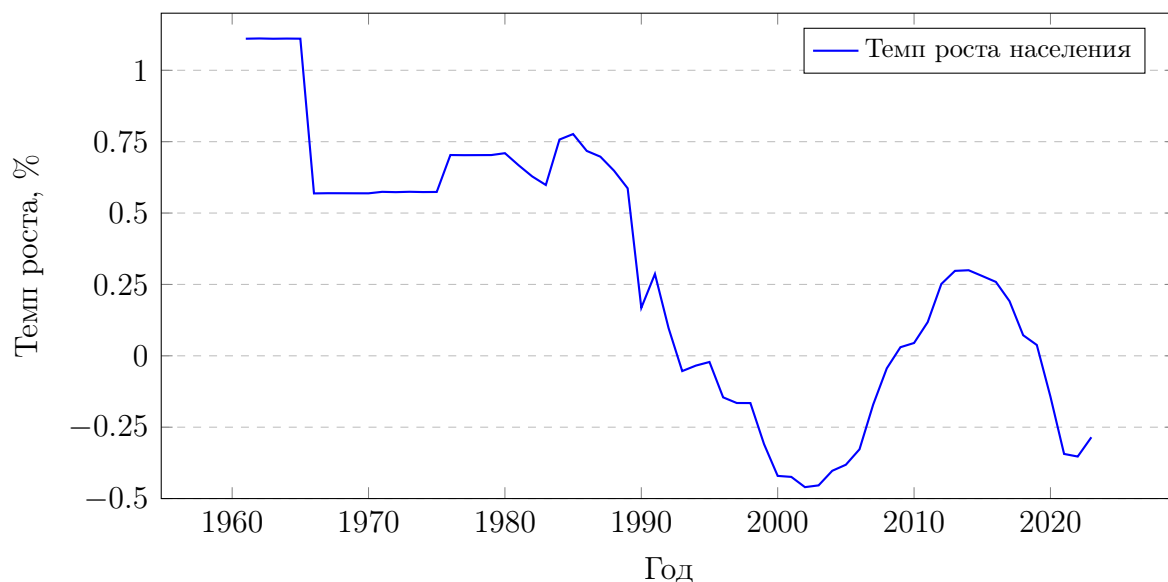
расширяет SAR за счёт учета пространственных лагов объясняющих переменных.

Пространственные модели позволяют точнее описывать межрегиональные взаимосвязи и учитывать распространение эффектов от соседних территорий.

Современные демографические тенденции в России характеризуются значительной региональной дифференциацией, что делает актуальным изучение факторов, влияющих на динамику численности населения. Анализ та-

ких факторов позволяет не только выявить ключевые закономерности, но и разработать адресные меры социально-экономической политики, направленные на стабилизацию демографической ситуации.

Рисунок - 1: Темп роста населения



В данном исследовании проводится эмпирический анализ данных по субъектам Российской Федерации за 2021–2023 годы, основанный на официальных статистических показателях Росстата. В качестве зависимой переменной рассматривается бинарный признак, отражающий наличие положительного прироста населения в регионе. Объясняющие переменные включают такие социально-экономические индикаторы, как доля городского населения, среднедушевые доходы, уровень безработицы и доля населения трудоспособного возраста.

Для оценки вероятности положительного прироста населения применяются две базовые модели бинарного выбора — logit и probit. Их сравнение позволяет проверить устойчивость результатов и выявить наиболее значимые факторы. Полученные выводы могут быть полезны для формирования региональной политики, направленной на стимулирование демографического роста и снижение дисбалансов в развитии территорий.

Опираясь на статистические данные Росстата, была сформирована информационная база для построения моделей:

Таблица 1 – Социально-экономические показатели регионов России за 2021 год

Регион	Прирост (%)	Городское население (%)	Доход (руб.)	Безработица (%)	Население трудоспособном возрасте (%)
Белгородская область	-0.6	65.3	35 511	4.2	56.9
Брянская область	-1.2	69.5	31 713	3.4	56.6
Владимирская область	-1.2	77.6	28 148	3.9	55.8
Воронежская область	-0.7	68.5	34 982	3.8	56.9
...	...	...	...	...	...
Магаданская область	-1.0	96.4	82 031	4.9	62.3
Сахалинская область	-0.7	82.5	66 211	5.2	59.3
Еврейская АО	-2.0	70.8	31 118	5.5	57.8
Чукотский АО	0.7	68.8	104 178	2.6	65.0

В качестве зависимой переменной выступает бинарный признак, принимающий значение 1 в случае положительного прироста численности населения в регионе и 0 — в случае нулевого или отрицательного прироста.

Были построены две базовые модели: logit и пробит. Обе модели оценивают вероятность положительного прироста населения в зависимости от объясняющих факторов.

Результаты обеих моделей выявили высокую значимость доли трудоспособного населения и доли городского населения. Первый показатель положительно влияет на вероятность прироста: с увеличением трудоспособного населения в регионе шансы на рост численности населения возрастают. При этом, высокая доля городского населения оказывает отрицательное влияние на вероятность прироста. Показатели дохода и безработицы не продемонстрировали статистической значимости, что требует дополнительного анализа и обсуждения. Сравнение logit- и probit-моделей показывает, что результаты качественно идентичны: знаки коэффициентов и состав значимых переменных совпадают. Различия заключаются лишь в масштабах коэффициентов, что обусловлено различной формой распределения ошибок в моделях. Выбор между ними не влияет на выводы.

На основе построенных logit- и probit-моделей можно сделать следующие выводы о влиянии факторов на вероятность положительного прироста населения:

- **Доля трудоспособного населения** является положительно значимым предиктором. Это может отражать более устойчивую социальную структуру в регионах с высоким средним возрастом, где выше уровень

оседлости и ниже миграционная подвижность.

- **Доля городского населения** оказывает отрицательное влияние на вероятность прироста. Высокая степень урбанизации может ассоциироваться с насыщенностью инфраструктуры, высокими расходами на проживание и ограниченными ресурсами для дальнейшего демографического роста.
- **Доход и уровень безработицы** не показали статистической значимости, что на первый взгляд противоречит интуитивным ожиданиям.

Несмотря на теоретические ожидания, среднедушевой доход и уровень безработицы оказались незначимыми. Возможные причины включают:

- **Недостаточную вариативность** по регионам: разброс значений может быть недостаточен для выявления устойчивой зависимости;
- **Временная задержка**: экономические факторы могут влиять на демографические показатели с опозданием, что не отражено в статической модели;
- **Нелинейные и пороговые эффекты**: возможно, влияние проявляется только при экстремальных значениях;
- **Опосредованное влияние**: доход и безработица могут воздействовать через другие переменные — миграцию, рождаемость, уровень образования и т.п.;
- **Особенности бинарной модели**: использование дискретной переменной снижает чувствительность модели к слабым связям.

Анализ показал, что социально-демографические характеристики — в первую очередь возраст и уровень урбанизации — оказывают статистически значимое влияние на вероятность прироста населения в регионах России. Экономические факторы, вопреки ожиданиям, не показали существенного влияния в рамках рассмотренных моделей. Это подчёркивает необходимость дальнейшего исследования с использованием расширенных пространственно-динамических моделей и дополнительного набора переменных.

ЗАКЛЮЧЕНИЕ

В настоящей работе рассмотрены пространственные регрессионные модели с дискретной зависимой переменной на примере анализа демографических данных по регионам России. В теоретической части были представлены базовые модели дискретного выбора (logit или probit), а также пространственные расширения классических регрессионных моделей (SAR, SEM, SDM), позволяющие учитывать межрегиональные взаимодействия.

Практическая часть была посвящена построению и анализу logit- или probit-моделей, объясняющих вероятность положительного прироста населения на основе социально-экономических факторов. В ходе моделирования установлено, что значимое влияние оказывают средний возраст населения и уровень урбанизации, тогда как доход и безработица не проявили статистически значимого эффекта.

Результаты анализа позволяют сделать следующие основные выводы:

- Простые модели дискретного выбора эффективно описывают демографическую динамику на региональном уровне;
- Социально-демографические характеристики оказывают более существенное влияние, чем макроэкономические показатели;
- Пространственные эффекты и временные лаги требуют дополнительного учета в дальнейшем анализе.

Полученные результаты представляют интерес для органов государственной власти при разработке региональной демографической и миграционной политики. В дальнейшем возможным направлением исследования может стать реализация пространственных logit- или probit-моделей, а также использование более широкого набора предикторов и динамических панельных данных.