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 ограничениями на зависимую переменную

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 412 группы
направления 01.03.02 — Прикладная математика и информатика

механико-математического факультета
Золотова Михаила Алексеевича

Научный руководитель
ст. преподаватель _____ А. Д. Луньков

Заведующий кафедрой
д. ф.-м. н., доцент _____ С. П. Сидоров

Саратов 2021

ВВЕДЕНИЕ

Актуальность темы. В современном мире зачастую встречаются случаи, когда нужно принимать управленческие решения относительно каких-либо географических объектов: от муниципальных районов до стран и континентов. В этих случаях очевидным является решение использовать географическую близость, как один из факторов, влияющих не принятие решения. Поэтому нельзя обойтись без использования пространственных регрессионных моделей.

Пространственные модели с ограничением на зависимую переменную все чаще используются в литературе по пространственной экономике. Это особенно верно для пространственной probit-модели, основанной на нормальном распределении, которой посвящена данная работа. Пространственная probit-модель может использоваться для объяснения эффектов взаимодействия между географическими единицами, когда зависимая переменная принимает форму бинарной переменной.

Целью бакалаврской работы является исследование методов оценивания параметров пространсвенной регрессионной модели с ограничениями на зависимую переменную, в частности пространственную probit-модель.

Объект исследования — панельные данные.

Предмет исследования — пространственные эконометрические модели на панельных данных, содержащих основные социально-экономические параметры регионов Российской Федерации.

Для достижения поставленных целей в работе необходимо решить следующие **задачи**:

— рассмотреть модели бинарного и множественного выбора, описать методы оценки их параметров;
— изучить модели с цензурированными выборками;
— рассмотреть панельные данные и основные модели с их использованием;
— рассмотреть пространственные регрессионные модели, в частности пространственную probit-модель;
— создать код, оценивающий параметры пространственной probit-модели;
— провести численный эксперимент, используя полученные оценки пара-

метров эконометрической модели.

Практическая значимость. Исследована зависимость размера средней начисленной пенсии от численности населения, процента населения пенсионного возраста, размера регионального ВВП и других. Модель калибрована на основе данных полученных с сайта Росстата и может быть полезна для определения зависимостей между уровнем начисленных пенсий в регионе и основными социально–экономическими параметрами региона, а также влияния соседних регионов с более высоким уровнем пенсии на регионы с более низким. Создан программный продукт и проанализированы по реальным современным социально-экономическим данным зависимости между вышеперечисленными параметрами. Результатам дана содержательная интерпретация.

Структура и содержание бакалаврской работы. Бакалаврская работа состоит из введения, семи разделов, заключения, списка использованных источников и пяти приложений. Общий объем работы составляет 50 страниц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обосновывается актуальность темы работы, формулируется цель работы и решаемые задачи, отмечается практическая значимость полученных результатов.

В **первом разделе** рассматриваются модели с ограничениями на зависимую переменную.

Модели бинарного выбора.

Ставится задача нахождения зависимости между принятием решения y («положительное» или «отрицательное», 1 или 0) и набором факторов $x = (x_1, ..., x_k)$, которые влияют на это решение.

Линейная модель вероятности.

Модель зависимости y_t от x_t можно записать в виде

$$y_t = x_t'\beta + \varepsilon_t, \qquad t = 1, ...n, \tag{1}$$

где t — номер наблюдения, $\beta = (\beta_1, ..., \beta_k)'$ — набор неизвестных коэффициентов, ε_t — случайная ошибка.

Или, учитывая, что y_t принимает значения 0 или 1, и принимая гипотезу о центрированности ошибок, модель можно переписать в виде

$$P(y_t = 1) = x_t'\beta. \tag{2}$$

Главным недостатком данной модели является то, что прогнозные значения, полученные после оценки коэффициентов методом наименьших квадратов, могут лежать вне области допустимых значений. Для устранения этого недостатка линейной модели далее в разделе рассматриваются **logit- и probit-модели**.

Полагаем, что существует некая функция $F(\cdot)$ с областью значения в отрезке [0,1], которая выражает зависимость $P(y_t = 1)$ от β:

$$P(y_t = 1) = F(x_t'\beta). \tag{3}$$

Рассмотрим наблюдаемую количественную переменную y_t^*, связанную с независимыми переменными x_t регрессионным уравнением

$$y_t^* = x_t'\beta + \varepsilon_t, \tag{4}$$

где $F(\cdot)$ — функция распределения ошибки.

Тогда не наблюдаемую величину можно будет определить следующим образом:

$$\begin{aligned} y_t &= 1, \quad \text{если } y_t^* \geq 0, \\ y_t &= 0, \quad \text{если } y_t^* < 0. \end{aligned} \tag{5}$$

В качестве функции $F(\cdot)$ можно использовать:

— функцию логистического распределения:

$$F(u) = \Lambda(u) = \frac{e^u}{1 + e^u},$$

и получить соответствующую ей logit-моделью;

— функцию стандартного нормального распределения:

$$F(u) = \Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{z^2}{2}} dz$$

и получить соответствующую ей probit-модель.

Для оценки коэффициентов β этих моделей используется метод максимального правдоподобия. Функция правдоподобия имеет следующий вид:

$$L = L(y_1, \ldots, y_n) = \prod_{y_t=0} (1 - F(x_t'\beta)) \prod_{y_t=1} F(x_t'\beta). \tag{6}$$

Проведя некоторые математические преобразования получаем систему нелинейных относительно β уравнений:

$$\frac{\partial l}{\partial \beta} = \sum_t \left(\frac{y_t p(x_t'\beta)}{F(x_t'\beta)} - \frac{(1 - y_t) p(x_t'\beta)}{1 - F(x_t'\beta)} \right) x_t = 0. \tag{7}$$

В общем случае нельзя аналитически найти решение данной системы, но для probit- и logit-моделей логарифмическая функция правдоподобия l является вогнутой по β фукнцией. Учитывая, что уравнение правдоподобия является необходимым условием локального экстремума, получаем, что решение системы дает оценку максимального правдоподобия коэффициентов β.

Модели множественного выбора.

Модели множественного выбора — это модели, в которых имеется более двух вариантов выбора. Их можно строить и изучать путем обобщения методов и подходов, используемых для моделей бинарного выбора, описанных в предыдущем разделе.

Предполагая, что ошибки ε_t независимы и имеют функцию распределения $F(x) = \exp(-e^{-x})$, а полезность u_{tj} зависит от наблюдаемых экзогенных (предопределенных) характеристик x_{tj} и неизвестных коэффициентов β, $u_{tj} = x_{tj}\beta$, получаем logit-модель множественного выбора:

$$P(y_t = j) = \frac{\exp(x_{tj}'\beta)}{\exp(x_{t1}'\beta) + \ldots + \exp(x_{tm}'\beta)}. \tag{8}$$

Одним существенным ограничением logit-модели множественного выбора является предположение о независимости полезностей по альтернативам. Это предположение невыполнимо в реальном мире, если альтернативы достаточно близки между собой.

Порядковые зависимые переменные

Далее рассматривается модель, в которой альтернативы упорядочены. Переменная y зависит от y^* следующим образом:

$$\begin{aligned} &y = 1, \quad \text{если } y^* \le c_1, \\ &y = 2, \quad \text{если } c_1 < y^* \le c_2, \\ &... \\ &y = m, \quad \text{если } c_{m-1} < y^* \le c_m, \end{aligned}$$

где $c_1, ..., c_m$ — некоторые фиксированные предельные уровни.

Предполагая независимости ошибок, функцию правдоподобия можно выразить следующим образом:

$$L = \prod_{j=1}^{m} \prod_{\{t:y_t=j\}} \left(F(c_j - x_t'\beta) - F(c_{j-1} - x_t'\beta)\right).$$

Максимизировав эту функцию, можно получить оценку параметров β и $c_j, j = 1, ..., m$.

Во **втором разделе** рассматриваются модели с цензурированными выборками.

Суть моделей с цензурированием состоит в том, что для части наблюдений известно не «истинное» значение зависимой переменной, а лишь ее усеченное значение, которое определяется уравнением цензурирования.

Tobit-модель.

Первая рассмотренная модель с цензурированием — tobit-модель, получена путем незначительной модификации ранее рассмотренной модели. Пусть наблюдаемая величина y удовлетворяет условию:

$$y_t = \begin{cases} y_t^*, & \text{если } y_t^* > 0, \\ 0, & \text{если } y_t^* \le 0, \end{cases} \tag{9}$$

а ненаблюдаемая — регрессионному уравнению

$$y_t^* = x_t'\beta + \varepsilon_t. \tag{10}$$

Для получения состоятельных и асимптотически несмещенных оценок коэффициентов β используют метод максимального правдоподобия. Функция правдоподобия имеет следующий вид:

$$L = \prod_{y_t=0} \left(1 - \Phi\left(\frac{x_t'\beta}{\sigma}\right)\right) \prod_{y_t>0} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(y_t - x_t'\beta)^2\right). \tag{11}$$

Первая группа сомножителей соответствует цензурированным наблюдениям, а вторая — всем остальным. В отличие от ранее рассмотренных probit- и logit-моделей, где параметры β и σ нельзя оценить по отдельности, здесь они в функции правдоподобия L «разделены» и каждый их них можно оценить.

Модель Хекмана.

Далее рассматривается модель, в которой вероятность принятия решения об участии и степень участия разделены и могут зависеть от разных факторов. Величина y_t^* удовлетворяет уравнению линейной регрессии (степень участия):

$$y_t^* = x_t'\beta + \varepsilon_t. \tag{12}$$

А решение об участии в мероприятии описывается моделью бинарного выбора:

$$g_t^* = z_t'\gamma + u_t, \tag{13}$$

$$\begin{aligned} g_t &= 1, \text{ если } g_t^* \geq 0, \\ g_t &= 0, \text{ если } g_t^* < 0, \end{aligned} \tag{14}$$

где z_t — экзогенные переменные, которые могут иметь общие компоненты с x_t; u_t — случайная ошибка. Наблюдения задаются в следующем виде:

$$\begin{aligned} &y_t = y_t^*, g_t = 1, \text{ если } g_t^* \geq 0, \\ &y_t \text{ не наблюдается, } g_t = 0, \text{ если } g_t^* < 0. \end{aligned} \tag{15}$$

Полученная модель называется моделью Хекмана. Очевидно, что при

$x_t = z_t,\ \beta = \gamma,\ \varepsilon_t = u_t$ получается обычную tobit-модель.

Оценить модель Хекмана можно с помощью метода максимального правдоподобия.

Третий раздел посвящен панельным данным.

Панельные данные состоят их повторных наблюдений одних и тех же выборочных единиц, которые осуществляются в последовательные периоды времени.

Основные модели.

За y_{it} обозначим зависимую переменную для исследуемого экономического объекта i в момент времени t, x_{it} — набор объясняющих переменных (вектор размерности k) и ε_{it} — ошибка, $i = 1, ..., n,\ t = 1, ..., T$. Обозначим

$$y_i = \begin{bmatrix} y_{i1} \\ \dots \\ y_{iT} \end{bmatrix},\ y = \begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix};$$

$$X_i = \begin{bmatrix} x'_{i1} \\ \dots \\ x'_{iT} \end{bmatrix},\ X = \begin{bmatrix} X_1 \\ \dots \\ X_n \end{bmatrix};$$

$$\varepsilon_i = \begin{bmatrix} \varepsilon_{i1} \\ \dots \\ \varepsilon_{iT} \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{bmatrix}.$$

Простейшая модель — модель линейной регрессии $y = X\beta + \varepsilon$, не учитывающая панельную структуру данных.

Одна из возможных реализаций модели, в которой благодаря панельным данным можно учитывать индивидуальные различия между экономическими объектами выглядит следующим образом:

$$y_{it} = \alpha_i + x'_{it}\beta + \varepsilon_{it}, \tag{16}$$

где α_i — индивидуальный эффект объекта i, не зависящий от времени t, регрессоры не содержат константу.

Предполагая, что в уравнении (16) α_i — неизвестные параметры, рас-

сматривается так называемая модель с фиксированным эффектом. Если же предположить, что $\alpha_i = \mu + u_i$, где μ — общий, фиксированный во времени параметр для всех объектов, а u_i — ошибки, некоррелированные с ε_{it} и некоррелированные при разных i, то полученная модель называется моделью со случайным эффектом.

Методы оценивания опираются на понижение размерности вектора неизвестных переменных (удаление среднего), на классический и обобщенный МНК.

В **четвертом разделе** рассмотрены пространственные модели.

Пространственные модели используются для исследования взаимосвязей между двумя географическими объектами.

Модель пространственного лага получается при добавления к обычной линейной модели (1) компоненты пространственной авторегрессии, которая моделирует пространственный лаг:

$$Y = X\beta + \rho WY + \varepsilon, \tag{17}$$

где Y — вектор N x 1, состоящий из значений зависимых переменных для каждого объекта, X — матрица N x K, состоящая из N измерений для K независимых переменных, β — вектор коэффициентов K x 1, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_N)'$ — вектор ошибок с многомерным нормальным распределением, нулевым средним и дисперсией $\sigma^2 I$, ρ — коэффициент регрессии, отражающий степень пространственной автокорреллляции. В качестве W обозначили известную N x N матрицу пространственных весов, которая показывает соседство исследуемых объектов. По диагонали W заполнена нулями.

Для оценки SAR зачастую используют метод максимального правдоподобия. Логарифмическая функция максимального правдоподобия имеет вид:

$$\begin{aligned} \ln L = &- \frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\sigma^2 + \ln|I - \rho W| - \\ &- \frac{1}{2\sigma^2}(Y - \rho WY - X\beta)^T(Y - \rho WY - X\beta). \end{aligned} \tag{18}$$

Получение оценок по данной функции правдоподобия бывает затруд-

нительным, поэтому на практике переходят к концентрированной функции максимального правдоподобия, используя условия первого порядка.

Далее в разделе рассматривается **модель пространственной ошибки** (SEM – Spatial Error Model):

$$\begin{aligned} Y &= \rho X\beta + u \\ u &= \lambda W u + \varepsilon, \end{aligned} \tag{19}$$

где λ — коэффициент пространственной автокорреляции.

Для оценки модели SEM используется метод максимального правдоподобия. Функция правдоподобия имеет вид:

$$\begin{aligned} \ln L = &- \frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\sigma^2 + \ln|I - \rho W| - \\ &- \frac{1}{2\sigma^2}(Y - X\beta)^T(I - \lambda W)^T(I - \lambda W)(Y - X\beta). \end{aligned} \tag{20}$$

После концентрации функции правдоподобия получим оценки:

$$\hat{\beta}_{ML} = \left((X - \lambda WX)^T(X - \lambda WX)\right)^{-1}(X - \lambda WX)^T(Y - -\lambda WY), \tag{21}$$

$$\hat{\sigma}^2_{ML} = (e - \lambda We)^T(e - \lambda We)/N, \tag{22}$$

где $e = Y - X\hat{\beta}$.

Пространственная probit-модель.

Обычная пространственная probit-модель имеет вид модели пространственной ошибки. В векторной форме модель можно записать в следующем виде:

$$Y^* = X\beta + \varepsilon,\ \varepsilon = \lambda W\varepsilon + v. \tag{23}$$

В этом случае ε — пространственно коррелированные ошибки с коэффициентом λ, а v имеют многомерное нормальное распределение с нулевым средним и дисперсией I.

Пятый раздел посвящен оценке параметров probit-модели.

Поскольку в пространственной probit-модели, в отличие от обычной (не

пространственной), компоненты ошибки ε_t зависят друг от друга, и функция правдоподобия будет являться N-мерным интегралом:

$$L(\beta, \lambda|Y) = \int_{Y^*} \frac{1}{\sqrt{(2\pi)^N |\Omega_\lambda|}} \exp\left(-\frac{1}{2}\varepsilon' \Omega_\lambda^{-1} \varepsilon\right) d\varepsilon. \tag{24}$$

Он показывает вероятность того, что получена выборка Y. Например, для $N = 3,\ Y = (0, 1, 0)'$, функция правдоподобия равна вероятности выполнения следующей системы:

$$\begin{cases} y_1^* < 0, \\ y_2^* \geq 0, \\ y_3^* < 0. \end{cases}$$

Одной из сложностей, возникающих при вычислении данного интеграла, является нахождение численными методами обратной матрицы $(I - \lambda W)$ для больший значений N. Большинство численных методов, используемых для нахождения обратной матрицы размерности N x N работают за $O(N^3)$, что очень затратно по времени для $N > 1000$.

В **шестом разделе** была рассмотрена расширенная probit-модель.

Предположим, что существует два состояния объекта — 0 и 1, y_i определяет состояние объекта $i, i = 1, \ldots, N$, и данный объект i переходит из состояния 0 в 1 в момент времени t_i $(t = 1, \ldots, T)$. Стоит задача определения зависимостей, которые объясняют переход объекта в состояние 1.

Можно сделать предположение, что объекты в состоянии 1, имеют отличное влияние на своих соседей от объектов, которые пока остаются в состоянии 0. Отразим это в модели, двумя различными коэффициентами регрессии ρ и δ для этих двух переменных. Тогда, учитывая сделанные предположения, расширенную probit-модель можно записать в следующем виде:

$$Y_t^{0*} = \rho W_t^{00} Y_t^{0*} + \delta W_t^{01} Z_t + X_t^0 \beta + v_t^0. \tag{25}$$

Параметры данной модели также оцениваются методом максимального правдоподобия.

В **седьмом разделе** был проведен вычислительный эксперимент.

В ходе работы написана программа с использованием пакета MatLab,

оценивающая параметры расширенной probit-модели на реальных данных. В качестве зависимой переменной y^* была выбрана средняя начивленная пенсия в регионе, а индикатор y принимал значение 1, когда размер пенсии в регионе превышал 12000 рублей.

Была собрана статистика по 82 регионам Российской Федерации (краев, областей, республик). Для исследования были выбраны основные социально-экономические показатели за 2013-2018 года: численность населения, размер регионального ВВП, среднедушевые потребительские расходы за месяц и другие. Информация была взята с сайта Росстата https://rosstat.gov.ru/.

В результаты были получены оценки параметров модели, приведенные в таблице 1.

Таблица 1 – Оценки пространственных probit-моделей

Переменная	Стандартная probit-модель	Расширенная probit-модель
ρ	0.7637*** (5.07)	0.7393*** (5.68)
δ	—	-39.0204 (-1.11)
Население	-2.2761* (-1.84)	-3.3322** (-2.03)
Население пенсионного возраста	0.0002 (1)	0.0002 (0.67)
ВРП	0.0106 (0.11)	0.0140 (0.14)
Потребительские расходы	0.0001*** (31.42)	0.0001*** (23.7)
Общая жил. площадь	0.0174 (0.39)	0.0421 (0.3)
$\log L$	-120.18	-18.85
Время работы	14 минут 4.7 секунд	4 минуты 44.9 секунд

Поскольку для оценки параметров обеих моделей были использованы одинаковые данные и выбраны одинаковые регрессоры, можно провести сравнение полученных оценок. Основываясь на результатах, приведенных в таблице 1, можно сделать выводы, что коэффициент пространственной авто-

корреляции ρ является положительным и значимым на уровне 1% в обеих моделях. В то же время в расширенной probit-модели параметр δ, показывающий уровень автокорреляции между объектами в состоянии 0 и объектами в состоянии 1, является отрицательным и незначимым даже на уровне 10%. Следовательно, данный результат можно проинтерпретировать следующим образом: после перехода в состояние 1 объекты оказывают незначительное влияние на объекты, которые все еще в состоянии 0.

Полученное преимущество во времени работы программы будет намного более значимым, когда, например, будут рассматриваться более длительные промежутки времени и большее количество определяющих переменных.

В **заключении** приведены результаты бакалаврской работы.

Основные результаты

1. Рассмотрены модели бинарного и множественного выбора, описаны методы оценки их параметров.

2. Изучены модели с цензурированными выборками и определены области их использования.

3. Рассмотрена структура панельных данных и основные модели с их использованием.

4. Рассмотрены пространственные регрессионные модели, в частности пространственная probit-модель.

5. Создана программа, позволяющая оценить параметры пространственной probit-модели для собранных данных по российским регионам.