МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

Анализ мер близости между регионами в пространственных регрессионых моделях

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

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

механико-математического факультета
Рубцова Олега Александровича

Научный руководитель
ст. преподаватель _____ А. Д. Луньков

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

Саратов 2020

ВВЕДЕНИЕ

Актуальность темы. Проблема качественного анализа разнородных статистических данных всегда была и продолжает оставаться актуальной. Одним из основных инструментов анализа данных является регрессионный анализ. Пространственная эконометрика - это подраздел эконометрики изучающий пространственное взаимодействие между географическими единицами. Этими единицами могут быть почтовые индексы, города, регионы, округа, в зависимости от характера исследования. Пространственные модели учитывают взаимосвязи между единицами наблюдения с помощью весовых матриц, пространственные и временные эффекты. Все это позволяет специфицировать более сложные гипотезы, включая индивидуальные эффекты, которые не могут быть рассмотрены в моделях перекрестных данных.

Целью бакалаврской работы является исследование методов оценивания параметров пространственных моделей применительно к российским региональным данным и некоторым социально-экономическим показателям.

Объект исследования – пространственные модели.

Предмет исследования – пространственные экономические модели.

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

- рассмотреть основные модели панельных данных, их гипотезы;
- рассмотреть динамические модели с панельными данными;
- рассмотреть модели бинарного выбора с панельными данными;
- описать метод обобщенных моментов для получения оценок параметров регрессионых моделей;
- рассмотреть основные пространственные модели;
- создать программный код, позволяющий оценить параметры пространственных моделей;
- проанализировать полученные результаты.

Практическая значимость. Исследована зависимость среднего дохода населения в регионах от показателей численности, занятости населения и количества предприятий. Модель построена на основе данных, полученных с портала www.gks.ru и может быть полезна для прогнозирования среднего дохода

в регионах. Создан программный продукт. Результатам дана содержательная интерпретация.

Структура и содержание бакалаврской работы. Работа состоит введения, трех разделов, заключения и списка использованных источников, содержащего 20 наименований. Общий объем работы составляет 40 страниц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обосновывается актуальность темы работы, описывается цель работы, ее практическая значимость.

В **первом** разделе рассматривается модель панельных данных.

Обозначения и основные модели.

Пусть y_{it} - зависимая переменная для экономической единицы i в момент времени t, x_{it} - набор объясняющих переменных (вектор размерности k) и ε_{it} - соответствующая ошибка, $i = 1, ..., n,\ t = 1, ..., T$. Обозначим

$$y_i = \begin{bmatrix} y_{i1} \\ \dots \\ y_{iT} \end{bmatrix},\ X_i = \begin{bmatrix} x'_{i1} \\ \dots \\ x'_{iT} \end{bmatrix},\ \varepsilon_i = \begin{bmatrix} \varepsilon_{i1} \\ \dots \\ \varepsilon_{iT} \end{bmatrix}.$$

Также обозначим

$$y = \begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix},\ X = \begin{bmatrix} X_1 \\ \dots \\ X_n \end{bmatrix},\ \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{bmatrix}.$$

Модель с фиксированных эффектом

Спецификация модели: $y_{it} = \alpha_i + x'_{it}\beta + \varepsilon_{it}$, где величина α_i выражает индивидуальный эффект объекта i, не зависящий от времени t, и регрессоры не содержат константу.

Предположим, что выполнены условия:

1) ошибки ε_{it} некоррелированы между собой по i и t, $E(\varepsilon_{it}) = 0,\ V(\varepsilon_{it}) = \sigma^2_\varepsilon$;
2) ошибки ε_{it} некоррелированы с регрессорами x_{js} для любых i, j, t, s.

Введем фиктивные переменные $d_{ij} = 1$, если $i = j$, и $d_{ij} = 0$, если $i \neq j$. И объединим их в одну матрицу:

$$D = \begin{bmatrix} \iota_T & 0 & \dots & 0 \\ 0 & \iota_T & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \iota_T \end{bmatrix} = I_n \otimes \iota_T,$$

где $\iota_T = [1, \dots, 1]'$ имеет размерность T, а I_n - единичная матрица размера n.

Тогда уравнение спецификацию можно переписать в следующем виде:

$$y = D\alpha + X\beta + \varepsilon, \tag{1}$$

где $\alpha = [\alpha_1, \dots, \alpha_n]'$.

Введем матрицу $M_D = I_{nT} - D(D'D)^{-1}D'$, которая осуществляет отклонения от индивидуальных средних. Тогда уравнение отклонениях от средних может быть записано так:

$$M_D y = M_D X\beta + M_D\varepsilon, \tag{2}$$

данное преобразование называется внутригрупповым преобразованием.

Применяя к (2) метод наименьших квадратов получим оценки

$$\hat{\beta} = (X'M_D X)^{-1} X M_D y. \tag{3}$$

Из формулы (3) вытекает оценка матрицы ковариаций оценки $\hat{\beta}_{FE}$:

$$V(\hat{\beta}_{FE}) = \sigma_\varepsilon^2 (X'M_D X)^{-1}.$$

Модель со случайным эффектом

Спецификация модели:

$$y_{it} = \mu + x'_{it}\beta + u_i + \varepsilon_{it}, \tag{4}$$

где μ - константа, а u_i - случайная ошибка, инвариантная по времени для каждой экономический единицы.

Будем считать, что выполнены условия:

1) ошибки ε_{it} некоррелированы между собой по i и по t, $E(\varepsilon_{it}) = 0,\ V(\varepsilon_{it}) = \sigma_\varepsilon^2$;
2) ошибки ε_{it} некоррелированы с регрессорами x_{js} для любых i, j, t, s;
3) ошибки u_i некоррелированы, $E(u_i) = 0,\ V(u_i) = \sigma_u^2$;
4) ошибки u_i некоррелированы с регрессорами x_{jt} для любых i, j, t;
5) ошибки u_i некоррелированы с регрессорами ε_{jt} для любых i, j, t.

Эту модель можно рассматривать как линейную модель, в которой ошибка $\omega_{it} = u_i\varepsilon_{it}$ имеет некоторую специальную структуру. Следуя введенным обозначениям, уравнение $y_{it} = \mu + x_{it}'\beta + u_i + \varepsilon_{it}$ можно переписать в виде

$$y = \mu\iota_{nT} + X\beta + \omega, \tag{5}$$

где ω - матрица размерности nT.

Матрица ковариаций ω имеет вид $\Sigma = E(\omega_i\omega_i') = \sigma_u^2\iota_T\iota_T' + \sigma_\varepsilon^2 I_T$. Тогда для объединенных наблюдений $\Omega = I_n \otimes \Sigma$.

Согласно обобщенному методу наименьших квадратов получаем

$$\begin{bmatrix}\hat{\mu}_{GLS}\\ \hat{\beta}_{GLS}\end{bmatrix} = \left(\begin{bmatrix}\iota_{nT}'\\ X'\end{bmatrix}(I_n \otimes \Sigma^{-1})\right)^{-1}\begin{bmatrix}\iota_{nT}'\\ X'\end{bmatrix}(I_n \otimes \Sigma^{-1})y. \tag{6}$$

Полученные оценки называются оценками со случайным эффектом: $\hat{\beta}_{GLS} = \hat{\beta}_{RE}$.

Преобразуем уравнение $y_i = \mu + x_{it}'\beta + u_i + \varepsilon_{it}$, взяв средние значения по времени для каждой экономической единицы

$$\overline{y}_{it} = \mu + \overline{x}_{it}'\beta + u_i + \overline{\varepsilon}_i.$$

Оценки, которые получаются применением к данному уравнению обычного метода наименьших квадратов, называются межгрупповыми: $\hat{\beta} = \hat{\beta}_B$. Эти оценки являются несмещенными и состоятельными при $n \to \infty$, не неэффективными. В последнем уравнении можно перейти уже к отклонениям от

глобальных средних и представить межжгрупповые оценки в виде:

$$\hat{\beta}_B = \Big(\sum_{i=1}^{n} (\overline{x}_i - \overline{x})(\overline{x}_i - \overline{x})' \Big)^{-1} \sum_{i=1}^{n} (\overline{x}_i - \overline{x})(\overline{y}_i - \overline{y})$$

В результате можно получить следующее представление для оценки со случайным эффектом:

$$\hat{\beta}_{RE} = W\hat{\beta}_B + (I_k - W)\hat{\beta}_{FE}, \tag{7}$$

где W - некоторая матрица, которую можно вычислить в явном виде и которая пропорциональная матрице, обратной матрице ковариаций оценки $\hat{\beta}_B$.

Таким образом, оценка со случайным эффектом является средневзвешенной внутри- и межгрупповой оценок.

Качество подгонки

Коэффициент детерминации R^2 интерпретируется как доля объясненной вариации зависимости переменной.

Для внутригрупповой регрессии коэффициент R^2 можно определить равенством

$$R^2_{within} = r^2(y_{it} - \overline{y}_i, \hat{y_{it}} - \hat{\overline{y}}_i),$$

где $\hat{y_{it}} - \hat{\overline{y}}_i = (x_{it} - \overline{x}_i)'\hat{\beta}_{FE}$ и $r^2(\cdot, \cdot)$ - выборочный коэффициент корреляции.

Аналогично можно определить данный коэффициент для межгрупповой регрессии:

$$R^2_{between} = r^2(\overline{y}_i, \hat{y}_i),$$

где $\hat{y}_i = \overline{x}_i'\hat{\beta}_B$.

Для обычной модели объединенный коэффициент детерминации

$$R^2_{overall} = r^2(y_{it}, \hat{y}_{it}),$$

где $\hat{y}_{it} = x_{it}'\hat{\beta}_{OLS}$.

Выбор модели

1. *Обычная модель против модели с фиксированным эффектом.* Тестирование может осуществлено с помощью обычного F-теста, проверяющего гипотезу $H_0 : \alpha_1 = \ldots = \alpha_n$ в модели с фиктивными переменными (1).
2. *Обычная модель против модели со случайным эффектом.* В этом случае требуется в модели со случайным эффектом (4) тестировать гипотезу $H_0 : \sigma_u^2 = 0$. Бреуш и Паган предложили тест множителей Лагранжа, основанный на следующей статистике:

$$LM = \frac{nT}{2(T-1)}\Big(\frac{e'DD'e}{e'e} - 1\Big)^2,$$

где e - остатки обычной регрессии. При гипотезе H_0 данная величина имеет распределение хи-квадрат с одной степенью свободы. Если $LM > \chi^2_\alpha(1)$, где $\chi^2_\alpha(1)$ – α-процентная точка распределения хи-квадрат с одной степенью свободы, то H_0 отвергается при уровне значимости α.
3. *Случайный эффект против фиксированного эффекта.* В этом случае необходимо проверить гипотезу $H_0 : Cov(\alpha_i, x_{jt}) = 0$, т.к. в модели со случайным эффектом предполагается, что индивидуальные эффекты не коррелируют с объясняющими переменными. Альтернативная гипотеза состоит в том, что эта ковариация отлична от нуля.

   Для проверки этой гипотезы можно использовать тест Хаусмана. Суть теста состоит в том, что при нулевой гипотезе оценки $\hat\beta_{RE}$ и $\hat\beta_{FE}$ не должны сильно отличаться, а если справедлива альтернативная гипотеза, то различие должно быть существенным. Можно показать, что при выполнении нулевой гипотезы из эффективности оценки $\hat\beta_{RE}$ следует асимптотическое равенство

$$V(\hat\beta_{FE} - \hat\beta_{RE}) = V(\hat\beta_{FE}) - V(\hat\beta_{RE}).$$

   Таким образом, статистика

$$\xi_{ll} = (\hat\beta_{FE} - \hat\beta_{RE})'(V(\hat\beta_{FE}) - V(\hat\beta_{RE}))^{-1}(\hat\beta_{FE} - \hat\beta_{RE})$$

   при нулевой гипотезе имеет асимптотически хи-квадрат распределение

с k степенями свободы, где $V(\hat{\beta}_{FE}),\ V(\hat{\beta}_{RE})$ - оценки соответствующих ковариционных матриц, а k - размерность вектора β.

Обобщенный метод моментов

Предположим, что модель включает переменные $y_i,\ x_i,\ z_i,\ i = 1, \ldots, n$, и пусть выполнены следующие равенства:

$$E(m_j(y_i, x_i, z_i, \theta)) = 0,\ j = 1, \ldots, l, \tag{8}$$

где $m_j(y_i, x_i, z_i, \theta)$ – некоторые известные скалярные функции, а θ – k-мерный вектор параметров. В моделях регрессии можно считать y_i зависимой переменной, x_i – набором регрессоров, z_i – инструментальными переменными.

Равенства (8) называют моментными тождествами или условиями ортогональности. Если ввести вектор-функцию

$$m(y_i, x_i, z_i, \theta) = (m_1(y_i, x_i, z_i, \theta), \ldots, m_l(y_i, x_i, z_i, \theta))',$$

то соотношения (8) можно записать в виде

$$E(m(y_i, x_i, z_i, \theta)) = 0. \tag{9}$$

Определим вектор функцию

$$g(y, X, Z, \theta) = \frac{1}{n} \sum_{i=1}^{n} m(y_i, x_i, z_i, \theta)$$

и запишем выборочный аналог равенства (9)

$$g(y, X, Z, \theta) = 0. \tag{10}$$

Далее для краткости $g(y, X, Z, \theta) = g(\theta)$. Если $k < l$, то модель не идентифицируема. Если $k = l$, то систему (10) можно разрешить относительно θ. Если $k > l$, то модель называется переопределенной.

Оценки обобщенного метода моментов параметров θ находятся путем

решения следующей задачи

$$g'(\theta)Sg(\theta) \to \min, \tag{11}$$

где S – некоторая симметричная положительно определенная матрица размерности $l \times l$.

Ясно, что разным весовым матрицам S соответствуют разные состоятельные оценки $\hat{\theta}_{GMM}$.

Обобщенный метод моментов обладает рядом преимуществ: для его использования не требуется знать распределение наблюдений, он работает при наличии гатероскедастичсности любого вида.

Динамические модели

Наиболее простую динамическую модель можно получить, добавляя в правую часть уравнения (**??**) лагированное значение зависимой переменной:

$$y_{it} = \alpha_i + x'_{it}\beta + \gamma y_{it-1} + \varepsilon_{it}. \tag{12}$$

В данном случае очевидно, что переменные y_{it-1} и α_i коррелированы независимо от природы индивидуального эффекта α_i.

Простейшая модель авторегрессии с панельными данными

Рассмотрим (12) без экзогенных переменных:

$$y_{it} = \alpha_i + \gamma y_{it-1} + \varepsilon_{it} \tag{13}$$

Состоятельные оценку параметра γ можно получить с помощью метода инструментальных переменных.

Модель с экзогенными переменными

Рассмотрим модель (12) с экзогенными переменными

$$y_{it} = \alpha_i + x'_{it}\beta + \gamma y_{it-1} + \varepsilon_{it}.$$

Переходя к первым разностям

$$y_{it} - y_{it-1} = (x_{it} - x_{it-1})'\beta + \gamma(y_{it-1} - y_{it-2}) + (\varepsilon_{it} - \varepsilon_{it-1}),$$

видим, что получается модель, аналогичная (13). Экзогенность регрессоров означает, что $E(x_{is}\Delta\varepsilon_{it}) = 0$ при всех $s,\ t$. Эти равенства можно рассматривать как моментные тождества.

Модели бинарного выбора с панельными данными

Модель бинарного выбора в случае панельных данных может быть описана следующим образом

$$y^*_{it} = x_{it}\beta + \alpha_i + \varepsilon_{it}, \tag{14}$$

где

$$y_{it} = 1,\ \text{если}\ y^*_{it} \geq 0,$$

$$y_{it} = 0,\ \text{если}\ y^*_{it} < 0,$$

где ошибки ε_{it} независимы по $i,\ t$ и одинаково распределены, а величины α_i отражают индивидуальные различия между объектами. Будем считать α_i неизвестными параметрами (модель с фиксированным эффектом). Для данной модели можно построить оценки максимального правдоподобия параметров $\alpha_i,\ i = 1, \ldots n,\ \beta$.

Пусть $f(y_{i1}, \ldots, y_{iT}|\alpha_i, \beta)$ - совместное распределение величин $y_{i1}, \ldots, y_{iT}$, зависящее от α_i, β. Предположим, что существует величина s_i, зависящая только от наблюдений, такая, что $f(y_{i1}, \ldots, y_{iT}|\alpha_i, \beta) = f(y_{i1}, \ldots, y_{iT}|s_i, \beta)$. Тогда максимизируя условную функцию правдоподобия

$$L^c = \prod_{i=1}^{n} f(y_{i1}, \ldots, y_{iT}|s_i, \beta),$$

можно получить состоятельные оценки параметров β, обладающие практически теми же свойствами, что и обычные оценки максимального правдоподобия.

Во **второй** части рассматриваются пространственная регрессионая мо-

дель с запаздыванием и фиксированным эффектом.

Третий раздел посвящен описанию эмпирической части.

В ходе работы были собраны данные о расстояниях от административных центров субъектов РФ до центров федеральных округов РФ и до ближайшего города с населением не меньше миллиона. Было взято 75 регионов РФ. Была исключена Республика Крым, т.к. для нее имеются данные только с 2015-го года. Посчитанные расстояния были занесены в таблицу Excel. Расстояния высчитывались с помощью сайта https://www.avtodispetcher.ru/distance.

Эти данные будут использоваться для построения весовых матриц в задачах пространственной эконометрики. Эти матрицы позволяют описывать взаимосвязь между единицами наблюдения в пространственных регрессионных моделях, в нашем случае регионами РФ.

В **заключении** приведены результаты работы.

Основные результаты работы

1. Рассмотрены основные модели панельных данных и их гипотезы.
2. Рассмотрены динамические модели с панельными данными, включая модели бинарного выбора.
3. Описан метод обобщенных моментов получения оценок параметров регрессии.
4. По регионам РФ был проведен анализ параметров построенной пространственной модели запаздывания с фиксированным эффектом. Рассмотрены два вида весовых матриц: расстояние до административного центра округа и до ближайшего города с населением не меньше миллиона.

   Создан программный код, позволяющий оценивать параметры пространственной модели запаздывания с фиксированным эффектом.