Министерство образования и науки Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

Динамические регрессионные модели для панельных данных

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 412 группы
направления 01.03.02 Прикладная математика и информатика

механико-математического факультета

Колесовой Светланы Валерьевны

Научный руководитель
ст. преп. А.Д. Луньков
должность, уч. степень, уч. звание подпись, дата инициалы, фамилия

Зав. кафедрой
д. ф.-м. н., доцент С.П. Сидоров
должность, уч. степень, уч. звание подпись, дата инициалы, фамилия

Саратов 2017

ВВЕДЕНИЕ

Актуальность темы исследования.

Одним из важнейших прикладных разделов математической статистики является регрессионный анализ, позволяющий моделировать недетерминированную взаимосвязь между некоторыми величинами. В 21 веке регрессионный анализ значительное внимание уделяет панельным данным. Совокупность панельных данных содержит повторные наблюдения для одних и тех же выборочных единиц, собранные за ряд временных периодов. Хотя панельные данные, как правило, собираются на микроэкономическом уровне, становится общепринятой практикой объединять индивидуальные временные ряды множества стран или множества отраслей промышленности и анализировать их одновременно. Применение повторных (для новых моментов времени) наблюдений, связанных с одними и теми же выборочными единицами, позволяет экономистам специфицировать и оценивать более сложные и более реалистические модели, чем те, что доступны для перекрёстных выборок или временных рядов. Совокупности панельных данных, тем не менее, очень часто страдают от пропущенных наблюдений. Даже если эти наблюдения отсутствуют просто из-за случайной потери информации, стандартная методика анализа должна быть с учетом таких пропусков скорректирована.

Панельные данные зачастую собираются по регионам, или по странам, как единицам наблюдения. Миграция, высокие технологии, торговые связи могут соединять экономические системы разных регионов, несмотря на территориальные границы. Игнорирование возможных пространственных взаимодействий при оценивании на основе региональных данных приводит к некорректным выводам в отношении величины и значимости влияния изучаемых факторов. В этом случае имеет место эффект пропущенных переменных, в результате полученные оценки будут смещенными и несостоятельными. Для учета пространственных связей в эмпирических задачах используется про-

странственная эконометрика. В основе пространственной методологии - использование пространственной весовой матрицы, элементы которой характеризуют наличие связей между регионами и интенсивность таких связей.

Целью работы является изучение современного инструментария пространственной эконометрики, а также создание программы, позволяющей оценивать параметры некоторых пространственных регрессионных моделей для данных по российским регионам.

Объект исследования - панельные данные, российские регионы, предмет исследования - эконометрические пространственные модели.

Актуальность работы связана с необходимостью учитывать при любом пространственном анализе экономических процессов взаимодействие географических единиц, или регионов, между собой и перспективы развития экономических процессов во времени с учетом предыстории. Все это и осуществляется с помощью современных статистических методов при построении динамических пространственных регрессионных моделей.

Актуальность определила выбор **темы** данной работы: «Динамические регрессионные модели для панельных данных».

Исследование имеет **практическую значимость**. Результаты оценивания параметров рассмотренных моделей могут быть полезны экономистам, изучающим конвергенцию регионов, их взаимосвязи. С помощью описанных методик можно выявить факторы, наиболее важные для объяснения величины того или иного показателя, характеризующего экономическое развитие региона (например, цены квадратного метра жилья или уровня безработицы).

Основное содержание работы

Выпускная квалификационная работа состоит из введения, 5 теоретических разделов, эмпирического раздела, заключения, списка использованных источников и приложений.

Введение содержит основные положения: цель, объект, практическую значимость задачи исследования.

Первый раздел «Элементы линейного регрессионного анализа» посвящен многомерной регрессионной модели.

Введем обозначения:

$y = (y_1, ..., y_n)'-$ вектор наблюдаемых значений зависимой переменной;

$\beta = (\beta_1, ..., \beta_k)'-$ вектор неизвестных коэффициентов;

$\epsilon = (\epsilon_1, ..., \epsilon_n)'-$ вектор ошибок;

$X = \begin{bmatrix} X_1 & \cdots & X_{1k} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{nk} \end{bmatrix} - n \times k$ матрица объясняющих переменных.

1. $y = X\beta + \epsilon-$ спецификация модели;

2. $X-$ детерминированная матрица, имеет ранг k;

3. $E(\epsilon) = 0, V(\epsilon) = E(\epsilon\epsilon') = \sigma^2 I_n$;

4. $\epsilon \sim N(0, \sigma^2 I_n)-$ дополнительное условие (при добавлении этого условия линейная регрессионная модель становится нормальной).

5. Условие независимости дисперсии ошибки от номера наблюдения называется свойством гомоскедастичностью и формулируется так:

$$E(\epsilon_t^2) = V(\epsilon_t) = \sigma^2, t = 1, ..., n.$$

Обратный случай соответствует гетероскедастичности.

6. Условие $E(\epsilon_t \epsilon_s) = 0, t \neq s$ указывает на некоррелированность ошибок для разных наблюдений. В случае, когда это условие не выполняется, говорят об автокорреляции ошибок.

В разделе рассмотрен **метод наименьших квадратов (ordinary least squares,**

OLS) - способ аппроксимации вектора y линейной функцией $f(\beta) = X\beta$, наилучшей в смысле минимизации функционала $F = (y - X\beta)'(y - X\beta)$.

Оценка вектора коэффициентов β методом наименьших квадратов для модели (1–4) имеет вид:

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y.$$

Ковариационная матрица МНК-оценки имеет вид:

$$V(\hat{\beta}_{OLS}) = \sigma^2(X'X)^{-1}.$$

Теорема Гаусса-Маркова.

В предположениях модели 1–3 оценка $\hat{\beta}_{OLS}$, полученная по методу наименьших квадратов, имеет наименьшую дисперсию в классе всех линейных несмещенных оценок.

Обобщенный метод наименьших квадратов (ОМНК).

Рассмотрим обобщенную регрессионную модель:

$$y = X\beta + \epsilon.$$

1. X - детерминированная матрица, имеет ранг k;

2. $E(\epsilon) = 0, V(\epsilon) = \Omega$, и матрица Ω положительно определена.

Теорема Айткена.

В классе линейных несмещенных оценок вектора β для обобщенной регрессионной модели оценка $\hat{\beta}_{GLS} = \hat{\beta}^* = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$ имеет наименьшую матрицу ковариаций.

Ковариационная матрица ОМНК-оценки:

$$V(\hat{\beta}^*) = (X'\Omega^{-1}X)^{-1}.$$

Во втором разделе «Анализ панельных данных» приведены общие сведения о панельных данных.

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

Обозначим $y_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{bmatrix}$, $X_i = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$, $\epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$

Простейшая модель для анализа панельных данных – это классическая линейная регрессионная модель:

$$y_{it} = x_{it}\beta + \epsilon_{it}$$

или в матричной форме, как и ранее:

$$y = X\beta + \epsilon$$

Такая форма записи не учитывает панельную структуру данных. При этом предполагается, что все ошибки ϵ_{it} некоррелированы между собой как по i, так и по t, и некоррелированы со всеми объясняющими переменными x_{it}. Эта модель носит название объединенной регрессионной. Ошибки имеют нулевое среднее и дисперсию $V(\epsilon_i) = \sigma^2$.

При выполнении сформулированных выше предположений OLS-оценки $\hat{\beta}_{OLS}$, по теореме Гаусса-Маркова, являются состоятельными и эффективными.

Панельные данные позволяют учитывать индивидуальные различия между объектами наблюдения. Для их выявления можно ввести в стандартную модель дополнительные составляющие следующим образом:

$$y_{it} = \alpha_i + x'_{it}\beta + \epsilon_{it},$$

где величина α_i характеризует индивидуальный эффект объекта i, не зависящий от времени t.

В зависимости от предположений относительно характера величины α_i рассматриваются две модели.

Модель с фиксированным эффектом: предполагается, что в уравнении α_i являются неизвестными параметрами, подлежащие оцениванию.

Формулы для оценки параметров:

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

$$\hat{\sigma}^2 = \frac{1}{nT - n - k}(y - X\hat{\beta});$$

$$\hat{\alpha}_i = \bar{y}_i - \bar{x}_i \hat{\beta}.$$

Здесь

$$M_D y = M_D X \beta + M_D \epsilon;$$

$$M_D = I_{nT} - D(D'D)^{-1} D'; D = I_n \otimes i_T.$$

Порой оценки этих параметров не представляет самостоятельного интереса, но их нельзя не учитывать.

Модель со случайным эффектом: предполагается, что $\alpha_i = \mu + u_i$, где μ – параметр, общий для всех единиц во все моменты времени, а u_i – ошибки, некоррелированные с ϵ_{it} и некоррелированные между собой при разных i. Для этой модели тоже описана методика оценивания.

В третьем разделе «Пространственные регрессионные модели» рассмотрены сведения о пространственной эконометрике вне контекста панельных данных. В двух предыдущих разделах все единицы наблюдения рассматриваются по отдельности. Географическая же, или любая другая степень близости объектов не принималась во внимание. Между тем такую зависимость нельзя не учитывать, например, при моделировании региональных экономических процессов (регион - единица наблюдения). В регрессионной модели возникают слагаемые, отвечающие за эффекты взаимодействия. Эти эффекты и являются предметом интереса в пространственной эконометрике.

Стандартное обоснование пространственных специфических эффектов: они заменяют в модели неучтенные, имеющие пространственную специфику, инвариантные во времени переменные, пропуск которых мог бы сместить оцен-

ки в перекрестной модели. Введем два важных символа для формализации вышеописанных эффектов:

$\rightarrow$ - взаимное влияние;

$\leftrightarrow$ - односторонняя причинно-следственная связь.

Стандартная линейная регрессионная модель, описанная в первой части, модифицируется (расширяется) для того, чтобы включить следующие слагаемые:

1) Эндогенный эффект взаимодействия (Wy):

- зависимая переменная y для единицы $A \leftrightarrow$ зависимая переменная y для единицы B;

Здесь W - неотрицательная матрица $n \times n$, описывающая степень пространственной взаимосвязи единиц в выборке и имеющая вид:

$$W = \begin{pmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nn} \end{pmatrix}.$$

Эндогенные эффекты взаимодействия (связаны с зависимой переменной y) - характеризуют склонность единиц наблюдения в некотором роде меняться в зависимости от поведения группы, к которой они принадлежат. Эти эффекты взаимодействия, как правило, рассматривают в качестве формальной спецификации для объяснения равновесия, возникающего в процессе того или иного пространственного или социально-экономического взаимодействия, благодаря которому значение зависимой переменной для одной единицы наблюдения совместно определяется значениями той же зависимой переменной для соседних единиц наблюдения. Наличие этого эффекта также называется пространственным лагом **(spatial lag model, SLM)**.

2) Внешние эффекты взаимодействия ($Wx\theta$):

- независимая переменная x для единицы $A \rightarrow$ зависимая переменная y для единицы B;

Внешние эффекты взаимодействия (связаны с независимой переменной x) связаны со склонностью единиц наблюдения в некотором роде меняться в зависимости от внешних характеристик всей группы (это k внешних объясняющих переменных, и таким образом, k внешних эффектов взаимодействия).

3) Эффект взаимодействия в ошибке ($\lambda W u$):

- ошибка u для единицы $A \leftrightarrow$ ошибка u для единицы B.

Этот эффект имеет следующую природу: люди(или иные объекты) в одной и той же группе склонны вести себя одинаково, потому что они имеют одни и те же индивидуальные характеристики или "живут"в одной и той же среде (эти эффекты "одинаковости"могут не наблюдаться). Эффекты взаимодействия в ошибке не требуют какой-либо теоретической модели для процесса пространственного или социального взаимодействия, как ранее. Они часто отражают ситуацию, при которой некоторые факторы, влияющие на зависимую переменную, но пропущенные в модели, пространственно автокоррелированы, когда ненаблюдаемые шоки носят пространственный характер. Наличие этого эффекта называется также пространственной зависимостью в ошибке, соответствующая регрессионная модель - **(spatial error model, SEM)**.

Линейная пространственная эконометрическая модель для перекрестных данных в векторных обозначениях с учетом всех вышеперечисленных эффектов имеет вид:

$$y = \rho W y + \alpha i_N + x\beta + Wx\theta + u; u = \lambda W u + \epsilon,$$

где $\alpha-$ свободный член, ρ - коэффициент пространственного лага, λ - коэффициент пространственной зависимости в ошибке, θ - вектор пространственного лага в регрессорах, u - пространственно автокоррелированная ошибка. Все слагаемые имеют размерность $n \times 1$. Общее количество эффектов взаимодействия в этой модели $k + 2$ (один - за счет зависимой переменной, еще один - за счет ошибки, остальные - за счет регрессоров).

Выбор спецификации W важен по следующим причинам:

1. Значения оценок параметров взаимодействия зависят от нее.

2. Уровень значимости каждого параметра взаимодействия зависит от нее.

3. Спецификация W должна соответствовать некоторым теоретическим предположениям. Однако, если обратиться к любой экономической теории для руководства к действию, то там, как правило, мало говорится о выборе W. Поэтому в эмпирических работах для принятия верного решения часто анализируется чувствительность результатов оценивания к виду W.

Элементы строки весовой матрицы характеризуют степень воздействия на определённую единицу со стороны других единиц, в то время как элементы столбца весовой матрицы дают представление о воздействии определённой единицы на все другие.

Пространственные весовые матрицы, чаще всего используемые в эмпирических исследованиях:

1. Бинарная матрица-индикатор смежности p-го порядка (если p=1, рассматриваются только соседи первого порядка, если p=2, рассматриваются только соседи первого и второго порядка, и так далее). При $p = 1$ бинарная матрица-индикатор имеет вид:

$$w_{ij} = \begin{cases} 1, & \text{если } j \text{ граничит с } i \\ 0, & \text{иначе} \end{cases}$$

2. Обратная матрица расстояний (иногда с некоторой границей пропускания - пороговой точкой):

$$w_{ij} = \frac{1}{(d_{ij})^{\nu}};$$

где ν - некоторое число.

Зачастую проводят нормализацию матрицы по строкам:

$$w_{ij}^{normalized} = \frac{w_{ij}}{\sum_{k=1}^{n} w_{ik}}.$$

Нормализация по строкам - стандартная процедура в пространственной эконометрике.

В разделе приведен обзор разновидностей линейной модели и методов оценки параметров.

В четвертом разделе «Модель пространственного лага и пространственной зависимости в ошибке» рассматриваются пространственные модели, но уже применительно к панельным данным. В отличие от третьего раздела здесь учитывается и время. Описанным выше эффектам, взятым по отдельности, соответствует своя регрессионная модель, две из них и рассматриваются в разделе. Если говорить о характере взаимодействия пространственных единиц, то модель может содержать пространственно лагированную зависимую переменную, а может наблюдаться процесс пространственной автокорреляции в ошибке. Эти модели - пространственный лаг (SLM) и модель пространственной зависимости ошибки (SEM) соответственно. Допустимо и сочетание этих эффектов. Согласно SLM значение зависимой переменной в одной единице определяется значениями той же переменной, наблюдаемой в соседних единицах, а также совокупностью наблюдаемых локальных характеристик данной единицы:

$$y_{it} = \rho \sum_{j=1}^{n} w_{ij} y_{jt} + x_{it}\beta + \alpha_i + \epsilon_{it}.$$

Согласно SEM зависимая переменная определяется набором наблюдаемых в данной единице локальных характеристик, а ошибки коррелированы в пространстве:

$$y_{it} = x_{it}\beta + \alpha_i + u_{it}; u_{it} = \lambda \sum_{j=1}^{n} w_{ij} u_{it} + \epsilon_{it}.$$

Для каждой из моделей дана полная спецификация. Эти модели можно оценить методом максимального правдоподобия, обобщенным методом моментов, алгоритмы описаны в разделе.

Пятый раздел носит название **«Динамические составляющие в регрессионных моделях».**

При построении большинства регрессионных моделей нельзя не учиты-

вать наличие "связи времен". Зачастую значение зависимой переменной в тот или иной момент времени связано со значением той же переменной или других факторов в предыдущие моменты, но классические панельные модели, как можно заметить, не учитывают этот факт.

Влияние только предыдущего момента (это простейшая ситуация) учитывается так: в правой части соотношения, описывающего регрессионную зависимость, появляются слагаемые вида τy_{it-1}. Эти модели можно оценить методом максимального правдоподобия, обобщенным методом моментов, марковскими цепями Монте-Карло. Динамическая составляющая может присутствовать как в перекрестных, так и в панельных моделях.

Динамическая пространственная модель для панельных данных выглядит так:

$$y_{it} = \tau y_{it-1} + \rho \sum_{j=1}^{n} w_{ij} y_{jt} + x_{it}\beta + \alpha_i + \epsilon_{it}.$$

Происходит переход к новым переменным:

$$y_{it}^* = y_{it} - \frac{1}{T} \sum_{t=1}^{T} y_{it},$$

тогда

$$y_{it}^* = \tau y_{it-1}^* + \rho \sum_{j=1}^{n} w_{ij} y_{jt}^* + x_{it}^*\beta + \alpha_i + \epsilon_{it}^*$$

Функция правдоподобия при некоторых вероятностных предположениях имеет вид:

$\log L = -\frac{nT}{2}\log(2\pi\sigma^2) + T\log|I_n - \rho W| - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it}^* - \tau y_{it-1}^* - \rho[\sum_{j=1}^{n} w_{ij}t_{jt}]^* - x_{it}^*\beta)^2$;

Оценке подлежат параметры $\tau, \beta, \sigma, \rho$. Оценим три параметра (все, кроме ρ):

$$\begin{bmatrix} \hat{\tau} \\ \hat{\beta} \end{bmatrix} = (\tilde{X}^{*'}\tilde{X}^*)^{-1}\tilde{X}^{*'}[Y^* - \rho(I_T \otimes W)Y^*],$$

$$\hat{\sigma}^2 = \frac{1}{nT}(Y^* - \rho(I_T \otimes W)Y^* - \tilde{X}^* \begin{bmatrix} \hat{\tau} & \hat{\beta}' \end{bmatrix}')'(Y^* - \rho(I_T \otimes W)Y^* - \tilde{X}^* \begin{bmatrix} \hat{\tau} & \hat{\beta}' \end{bmatrix}'),$$

здесь

$$\tilde{X}^* = \left[Y^*_{-1} \quad X^* \right].$$

В итоге концентрированная относительно ρ функция правдоподобия имеет вид:

$$\log L_C = C - \frac{nT}{2} \log[(e_0 - \rho e_1)'(e_0 - \rho e_1)] + T \log |I_n - \rho W|,$$

e_0 и e_1 - остатки некоторых вспомогательных регрессий, C - константа. Функция выпукла, ее максимум ищется численно.

Для эмпирической части «Оценка параметров регрессионной модели для экономических показателей российских регионов. Описание расчетов и программного кода» с помощью встроенного языка математической лаборатории Matlab написана программа, позволяющая оценивать параметры для динамической модели пространственного лага применительно к панельным данным. Код, результаты расчётов, фрагмент матрицы расстояний, на базе которой построена весовая матрица, содержатся в приложениях. В качестве элементов весовой матрицы берутся индикаторы смежности и величины, обратные к расстояниям.

Рассматривается построение регрессионной модели для данных по регионам России. Данные получены с сайта gks.ru. Модель определяют следующие переменные: зависимая переменная – уровень обеспеченности жильем, в числе регрессоров - уровень безработицы, плотность населения, ввод жилья.

Для этих данных оценены коэффициенты при регрессорах, дисперсия ошибки, коэффициент пространственного лага, коэффициент временного лага. Максимизация функции правдоподобия проводится численно. Данные для регрессионной модели могут подаваться на вход и в виде искусственно сгенерированного набора.

ЗАКЛЮЧЕНИЕ

В работе рассмотрены основные регрессионные модели для панельных данных. Рассмотрены пространственные и динамические модификации этих моделей. Основное внимание уделено модели пространственного лага. Для моделей описаны методы оценки параметров. Написаны программы, позволяющие оценивать параметры для реальных и сгенерированных данных. Результатам дана содержательная интерпретация.