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

ARCH-модели для панельных данных

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 412 группы
направления 01.03.02 Прикладная математика и информатика
Медведкова Андрея Алексеевича

Научный руководитель
ст. преп. _____ А.Д.Луньков

Зав. кафедрой
д.ф-м.н, профессор _____ С.П. Сидоров

Саратов, 2018

ВВЕДЕНИЕ

Актуальность темы. В современной жизни стали актуальными вопросы, связанные с использованием регрессионных моделей. Есть три базовых вида регрессионных моделей: модели перекрестных данных, модели временных рядов и модели панельных данных. Экономические, социальные и географические единицы могут изучаться в контексте их эволюции во времени. ARCH-модель описывает регрессионные модели временных рядов с условно гетероскедастичной ошибкой. Она была разработана сугубо из практических соображений. При описании колебания акций были выявлены временные кластеры и построены классификации случайных процессов, проходящих с акциями. В современной эконометрике существует множество модификаций моделей ARCH и GARCH, некоторые применительно к панельным данным. Практически ни одно современное исследование практического характера в области финансовых современных рядов не обходится без ARCH и GARCH-моделей.

Особенностью современных регрессионных моделей является наличие следующих составляющих: учет взаимосвязи между единицами наблюдения с помощью весовых матриц, учет как пространственных, так и временных эффектов, наличие обратных связей, учитываемых с помощью системы одновременных уравнений, наличие режимов переключения между видами моделей. Первой из упомянутых составляющих уделено внимание в этой работе.

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

Объект исследования — временные ряды.

Предмет исследования — пространственно-временные эконометрические модели для роста ВВП, плотности населения, прироста населения и доли городского населения.

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

— рассмотреть модель парной регрессии и основные гипотезы, связанные с этой моделью;

— определить основные гипотезы, лежащие в основе модели множественной регрессии, описать методику построения оценок её параметров;

— рассмотреть основные регрессионные модели для панельных данных;

— изучить понятие стационарности временных рядов и смежные вопросы;

— рассмотреть методы приведения ряда к стационарному;

— рассмотреть ARCH и GARCH-модели, в частности DCC-GARCH – пространственную модификацию GARCH;

— создать код, позволяющий программно оценить параметры ARCH-модели;

— провести анализ полученных результатов. Помимо прочего, будет создана программа, позволяющая оценивать параметры ARCH-моделей для искусственно сгенерированных данных.

Практическая значимость. Исследована зависимость индекса ВВП в регионах от того же показателя в географически или экономически близких регионах, доли городского населения, миграционного прироста внутри региона и плотности населения. Модель калибрована на основе данных, полученных с портала www.gks.ru и может быть полезна для прогнозирования внутреннего валового продукта региона, построения классификации регионов с учетом ВВП, выявления специфических эффектов каждого региона применительно к ВВП. Создан программный продукт и проанализированы по реальным современным социально-экономическим данным зависимости между вышеперечисленными показателями. Результатам дана содержательная интерпретация.

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

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

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

В **первом** разделе рассматривается модель парной регрессии.

Подгонка кривой.

Ставится задача подобрать («подогнать») функцию $Y = f(X)$ из параметрического семейства функций $f(X, \beta)$, «наилучшим» способом описывающую зависимость Y от X.

В качестве меры отклонения функции $f(X, \beta)$ от набора наблюдений можно взять:

1. сумму квадратов отклонений $F = \sum_{t=1}^{n} \left(Y_t - f\left(X_t, \beta\right)\right)^2$,

2. сумму модулей отклонений $F = \sum_{t=1}^{n} \left|Y_t - f\left(X_t, \beta\right)\right|$,

3. $F = \sum_{t=1}^{n} g\left(Y_t - f\left(X_t, \beta\right)\right)$, где g – любое преобразование отклонения $Y_t - f\left(X_t, \beta\right)$, входящего в функционал F.

Линейная регрессионная модель с двумя переменными.

Модель зависимости Y_t от X_t можно записать в виде

$$Y_t = a + bX_t + \varepsilon_t, \quad t = 1, \ldots, n,$$

где X_t — неслучайная (детерминированная) величина, а Y_t, ε_t — случайные величины. Y_t называется объясняемой (зависимой) переменной, а X_t — объясняющей (независимой) переменной или регрессором. Уравнение, приведенное выше, также называется регрессионным уравнением.

Основные гипотезы

1. $Y_t = a + bX_t + \varepsilon_t, t = 1, \ldots, n,$ — спецификация модели.

2. X_t — детерминированная величина; вектор $(X_1, \ldots, X_t)'$ не коллинеарен вектору $r = (1, ..., 1)'$.

3. $\mathrm{E}\varepsilon_t = 0$, $\mathrm{E}(\varepsilon_t^2) = \mathrm{V}(\varepsilon_t) = \sigma^2$— не зависит от t.

4. $\mathrm{E}(\varepsilon_t \varepsilon_s) = 0$ при $t \neq s$, некоррелированность ошибок для разных наблюдений.

   Часто добавляется условие:

5. Ошибки $\varepsilon_t, t = 1, \ldots, n$, имеют совместное нормальное распределение: $\varepsilon_t \sim \mathrm{N}(0, \sigma^2)$.

В этом случае модель называется нормальной линейной регрессионной.

Условие независимости дисперсии ошибки от номера наблюдения (от регрессора X_t): $\mathrm{E}(\varepsilon_t^2) = \mathrm{V}(\varepsilon_t) = \sigma^2, t = 1, \ldots, n$, называется *гомоскедастичностью.*

Условие $\mathrm{E}(\varepsilon_t \varepsilon_s) = 0$, $t \neq s$ указывает на некоррелированность ошибок для разных наблюдений. В случае, когда это условие не выполняется, говорят об *автокорреляции ошибок.*

Далее в разделе рассматриваются методы оценивания параметров a, b, σ^2:

1. с помощью метода наименьших квадратов в предположении теоремы Гаусса-Маркова.

$$\hat{b} = \frac{n \sum X_t Y_t - (\sum X_t)(\sum Y_t)}{n \sum X_t^2 - (\sum X_t)^2},$$

$$\hat{a} = \frac{1}{n} \sum Y_t - \frac{1}{n} \sum X_t \hat{b} = \overline{Y} - \overline{X} \hat{b}.$$

А также оценка σ^2:

$$\widehat{\sigma}^2 = \frac{1}{n-2} \sum e^2 t$$

2. с помощью метода максимального правдоподобия:

$$\hat{b}_{ML} = \frac{\sum x_t y_t}{\sum x^2 t}; \qquad \hat{a}_{ML} = \overline{Y} - \hat{b}_{ML} \overline{X}; \qquad \widehat{\sigma}^2_{ML} = \frac{1}{n} \sum e_t^2.$$

Во **втором** разделе рассмотрена модель множественной регрессии.

Она является естественным обобщением модели с двумя переменными:

$$y_t = \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + \varepsilon_t, \qquad t = 1, \ldots, \ n,$$

где x_{tp} — значения регрессора x_p в наблюдении t, а $x_{t1} = 1, t = 1, \ldots, n$.

Основные гипотезы

Гипотезы, лежащие в основе модели множественной регрессии, являются естественным обобщением модели парной регрессии:

1. $y_t = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + \varepsilon_t, \quad t = 1, \ldots, \ n$,— спецификация модели.

2. $x_{t1}, \ldots, \ x_{tk}$-— детерминированные величины. Векторы $x_s = (x_{1s}, \ldots, x_{ns})'$, $s = 1, ..., k$ линейно независимы в R^n.

3 – 5 совпадают с гипотезами парной регрессионной модели.

В этом случае модель называется *нормальной линейной регрессионной.*

Далее в разделе оцениваются следующие параметры:

$$\widehat{\beta}_{OLS} = (X'X)^{-1}X'y - \text{методом наименьших квадратов,}$$

$$\widehat{\sigma}^2 = \frac{e'e}{n-k} = \frac{\sum e_t^2}{n-k}$$

А также проверяется гипотеза *линейного ограничения общего вида* $H_0 : H\beta = r$, с помощью следующей статистики:

$$F = \frac{(H\widehat{\beta} - r)'(H(X'X)^{-1}H')^{-1}(H\widehat{\beta} - r)/q}{e'e/(n-k)} \sim F(q, n-k).$$

Третий раздел посвящен панельным данным.

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

Обозначение и основные модели.

Пусть y_{it} — зависимая переменная для экономической единицы i в момент времени t, x_{it} — набор объясняющих (независимых) переменных (вектор размерности k) и ε_{it} — соответствующая ошибка, $i = 1, ..., n,\ t = 1, ..., T$. При переходе к векторам используются обозначения

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

Вводятся также "объединенные" наблюдения и ошибки:

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix}, \quad X = \begin{bmatrix} x'_1 \\ \vdots \\ x'_T \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_T \end{bmatrix}.$$

(Здесь $y, \varepsilon - nT \times 1$ векторы, $X - nT \times k$ матрица.)

Простейшая модель — это обычная линейная модель регрессии

$$y = X\beta + \varepsilon,$$

которая, по сути, не учитывает панельную структуру данных. Вместе с тем предполагается, что все ошибки ε_{it} некоррелированы между собой как по t, так и по i, и также некоррелированы со всеми объясняющими переменными x_{it}. Эта модель именуется объединенной моделью регрессии *(pooled model)*. При выполнении изложенных выше предположений обычные МНК-оценки $\hat{\beta}_{OLS}$ являются состоятельными и эффективными.

Панельные данные позволяют учитывать индивидуальные различия между экономическими единицами. Одна из возможных реализаций этой идеи выглядит следующим образом:

$$y_{it} = \alpha_i + x'_{it}\beta + \varepsilon_{it}, \tag{1}$$

где величина α_i выражает индивидуальный эффект объекта i, который не зависит от времени t, а регрессоры x_{it} не содержат константу.

Исходя из предположений относительно характера величины α_i рассматриваются две модели.

Модель с фиксированным эффектом (fixed effect model): предполагается, что в соотношении (1) величины α_i являются неизвестными параметрами.

Модель со случайным эффектом (random effect model): предполагается, что в соотношении (1) $\alpha_i = \mu + u_i$, где μ — параметр, общий для всех единиц во все моменты времени, а u_i — ошибки, некоррелированные с ε_{it} и некоррелированные при разных i.

Методы оценивания опираются на понижение размерности вектора неизвестных переменных (удаление среднего), на классический и обобщенный МНК. В общем случае оцениваются вектор коэффициентов β, α_i, дисперсия ошибки, дисперсия эффектов (в предположении случайного эффекта).

В **четвертом** разделе рассмотрена стационарность временных рядов и смежные вопросы.

Под *временным рядом* (*динамическим рядом*, или *рядом динамики*) в экономике подразумевается последовательность наблюдений некоторого признака (случайной величины) Y в последовательные моменты времени.

Важнейшей классической задачей при исследовании экономических временных рядов является выявление и статистическая оценка основной тенденции развития изучаемого процесса и отклонений от нее.

Стационарность

Ряд y_t называется *строго стационарным (strictly stationary)* или *стационарным в узком смысле*, если совместное распределение m наблюдений $y_{t_1}, y_{t_2}, ..., y_{t_m}$ не зависит от сдвига по времени, то есть совпадает с распределением $y_{t_1+t}, y_{t_2+t}, ..., y_{t_m+t}$ для любых m, $t, t_1, ..., t_m$. Так как для исследователя обычно представляют интерес средние значения и ковариации, а не все распределение, поэтому часто используется понятие *слабой стационарности (weak stationarity)* или *стационарности в широком смысле*, которое состоит в том, что среднее, дисперсия и ковариации y_t не зависят от момента времени t:

$$E(y_t) = \mu < \infty, \quad V(y_t) = \gamma_0, \quad Cov(y_t, y_{t-k}) = \gamma_k.$$

Из строгой стационарности следует слабая стационарность (при условии конечности первого и второго моментов распределения).

Вводится понятие автокорреляционной функции (*autocorrelation function*), ACF:

$$\rho_k = \frac{Cov(y_t, y_{t-k})}{V(y_t)} = \frac{\gamma_k}{\gamma_0}.$$

При этом $\rho_0 = 1$, а $|\rho_k| \leqslant 1$.

В **пятом** разделе рассмотрены методы приведения ряда к стационарному.

В этом разделе обсуждаются модели временных рядов в узком смысле, т. е. модели, объясняющие поведение временного ряда, исходя только лишь из его значений в предыдущие моменты времени.

Статистические свойства стационарных и нестационарных временных рядов существенно отличаются, и для их моделирования должны применяться различные методы. Большое внимание уделяется именно моделям стационарных временных рядов, так как многие временные ряды могут быть приведены к стационарному ряду после операций выделения тренда, сезонной компоненты или взятия разности.

Тренд

Рассматривается следующий временной ряд:

$$y_t = \alpha + \beta t + \varepsilon_t. \tag{2}$$

Здесь ряд y_t представлен в виде композиции детерминированной составляющей $\alpha+\beta t$ (*линейный тренд*) и случайной составляющей ε_t, которая является стационарным временным рядом с нулевым средним. Часто встречаются дру-

гие примеры тренда: квадратичный, $\alpha + \beta t + \gamma t^2$; экспоненциальный $\alpha e^{\beta t}$ и т. п.

Для того чтобы выделить тренд в модели (2) (и ей подобных), можно применить обычную технику оценивания параметров регрессионных уравнений, считая t независимой переменной. После этого получают ряд остатков, для описания которого можно будет применить модели стационарных временных рядов.

Сезонность

В экономических данных часто встречается сезонная компонента. Например, в квартальных данных может наблюдаться сезонная компонента с периодом 4:

$$y_t = S(t) + \varepsilon_t, \qquad S(t+4) \equiv S(t). \tag{3}$$

В этом случае ряд y_t представлен в виде композиции периодической детерминированной составляющей $S(t)$ (*сезонная компонента*) и случайной составляющей являющейся стационарным временным рядом с нулевым средним. Сезонную компоненту $S(t)$ можно представить в виде $S(t) = \beta_1 d_{1t} + \beta_2 d_{2t} + \beta_3 d_{3t} + \beta_4 d_{4t}$, где d_i — фиктивные (бинарные) переменные для кварталов. Для выделения сезонной компоненты можно применить методы оценивания параметров регрессий к соотношению:

$$y_t = \beta_1 d_{1t} + \beta_2 d_{2t} + \beta_3 d_{3t} + \beta_4 d_{4t} + \varepsilon_t. \tag{4}$$

Как и в случае выделения тренда, методы моделирования стационарных временных рядов применяются далее к ряду остатков регрессии (4).

Взятие последовательной разности

Взятие последовательной разности также приводит к стационарному процессу ряд (2) с линейным трендом:

$$\Delta y_t = \beta + u_t, \qquad u_t = \Delta\varepsilon_t = \varepsilon_t - \varepsilon_{t-1}.$$

В случае квадратичного тренда $\alpha + \beta t + \gamma t^2$ взятие первой разности не приводит к стационарному ряду, но если взять вторую разность $\Delta^2 y_t = \Delta(\Delta y_t) = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2})$, то справедливо соотношение:

$$\Delta y_t = \beta + \gamma(2t-1) + \Delta\varepsilon_t,$$
$$\Delta^2 y_t = 2\gamma + \Delta^2\varepsilon_t,$$

и $\triangle^2 y_t$ уже является стационарным временным рядом.

В случае наличия сезонной компоненты (3) устранить последнюю можно при помощи оператора взятия *сезонной* последовательной разности $\Delta_4 y_t = (1-L^4)y_t = y_t - y_{t-4}$.

Таким образом, применяя выделение тренда, сезонности и/или оператор последовательной (и сезонной) разности, часто можно получить из исходного временного ряда стационарный.

Шестой раздел посвящен ARCH и GARCH-моделям.

Суть ARCH-модели состоит в следующем. Предположим, что имеется регрессия временного ряда y_t на другие временные ряды (все ряды предполагаются стационарными):

$$y_t = x_t'\beta + u_t. \tag{5}$$

Из эмпирических наблюдений за поведением таких рядов, как процентные ставки, обменные курсы и т.п., было замечено, что наблюдения с большими и малыми отклонениями от средних имеют тенденцию к образованию кластеров. То есть периоды «спокойного» и «возмущенного» состояний рынка чередуются.

Далее рассматривается способ моделирования этого явления. Пусть $\sigma_t^2 = V(u_t|u_{t-1},...,u_{t-p}) = E(u_t^2|u_{t-1},...,u_{t-p})$ — условная дисперсия ошибок u_t. Эффект «кластеризации» возмущений можно объяснить следующей моделью зависимости условной дисперсии ошибок u_t от предыстории:

$$\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + ... + \alpha_p u_{t-p}^2. \tag{6}$$

Процесс (5)–(6) называется *авторегрессионной условно гетероскедастичной* моделью порядка p *(AutoRegressive, Conditional Heteroscedastic)*, ARCH(p).

Также существует более общая спецификация модели для уравнения условной дисперсии ошибок (6):

$$\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \ldots + \alpha_p u_{t-p}^2 + \gamma_1 \sigma_{t-1}^2 + \ldots + \gamma_q \sigma_{t-q}^2. \quad (7)$$

Такая модель (5) – (7) называется *обобщенной авторегрессионной условно гетероскедастичной* порядка *p, q (Generalized Auto-Regressive, Conditional Heteroscedastic)*, GARCH(*p, q*).

Седьмой раздел посвящен описанию эмпирической части.

В ходе работы были собраны данные по расстояниям между административными центрами субъектов Российской Федерации. Было взято восемьдесят регионов РФ (краев, областей, республик). Были исключены Чукотский автономный округ и Камчатский край т.к. для них невозможно рассчитать расстояние по дорогам. Посчитаны расстояния между ними и занесены в таблицу Excel. Расстояния высчитывались несколькими способами:

1. с помощью сайта https://www.avtodispetcher.ru/, на котором можно рассчитать расстояние не только по дорогам, но и по прямой;

2. посредством формулы расстояний по большой дуге (Great-circle distance):

$$d = r\Delta\sigma, \quad (8)$$

где $r \approx 6371$ — средний радиус Земли (в км), а $\Delta\sigma$ — угловая разница, которая высчитывается по модифицированной формуле гаверсинусов:

$$\Delta\sigma = \arctan \frac{\sqrt{(\cos\phi_2 \sin(\Delta\lambda))^2 + (\cos\phi_1 \sin\phi_2 - \sin\phi_1 \cos\phi_2 \cos(\Delta\lambda))^2}}{\sin\phi_1 \sin\phi_2 + \cos\phi_1 \cos\phi_2 \cos(\Delta\lambda)}$$

где λ_1, λ_2 — долгота и ϕ_1, ϕ_2 — широта двух точек в радианах, $\Delta\lambda$ — разность координат по долготе.

Формула (8) помогает высчитать расстояние между двумя точками на Земле используя их географические координаты, а именно долготы и широты. Кратчайшим расстоянием между ними является длина дуги круга, проведенного на сфере по этим двум точкам. Поскольку в расчете участвует радиус, а у Земли, как у не совсем правильной сферы, он разный, на северном полюсе (6356.752 км), а на экваторе (6378.137 км), то в расчетах берется среднее значение (6371.008 км), что дает погрешность около 0.5%.

Для полученных расстояний была построена обратная матрица, а также вычислены коэффициенты корреляции для средней выборки (30 городов),

что бы узнать зависимость изменения расстояний:

1. по формуле — прямое = 0,988426115
2. прямое — по дорогам = 0,979964874
3. по формуле — по дорогам = 0,951021915.

В результате видно, что коэффициент корреляции близок к единице, что свидетельствует о том, что расстояния коррелированы.

Эти данные будут использоваться для построения весовых матриц в задачах пространственной эконометрики. Такие матрицы играют важную роль в пространственных регрессионных моделях. Они позволяют описывать взаимосвязь между единицами наблюдения, в нашем случае российскими регионами.

В **заключении** приведены результаты бакалаврской работы.

Основные результаты

1. Рассмотрена модель парной регрессии и основные гипотезы связанные с линейной регрессионной моделью.

2. Определены основные гипотезы, лежащие в основе модели множественной регрессии, описана методика построения оценок их параметров.

3. Рассмотрены основные регрессионные модели для панельных данных.

4. Определены основные понятия, связанные с временными рядами, изучена стационарность временных рядов и смежные вопросы, такие как: единичные корни, мнимая регрессия и коинтеграция.

5. Рассмотрены методы приведения ряда к стационарному.

6. Изучены ARCH и GARCH-модели.

7. По российским регионам был проведен предварительный анализ потенциальных весовых матриц. Было рассмотрено три вида весовых матриц: расстояние напрямую, по большой дуге и по дорогам. Корреляционный анализ установил значительную линейную зависимость между этими расстояниями. Это позволяет говорить о том, что можно пользоваться любой из этих матриц хотя бы в случае недостатка информации.

Создана программа, позволяющая оценивать параметры ARCH-моделей для искусственно сгенерированных данных.