ОСНОВНЫЕ ПРЕДПОСЫЛКИ РЕГРЕССИОННОГО АНАЛИЗА

ВОПРОСЫ

1. Основные предпосылки регрессионного анализа. Теорема Гаусса-Маркова.

2. Гетероскедастичность. Тест Голдфелда–Квандта.

3. Автокорреляция. Критерий Дарбина – Уотсона.

1. ОСНОВНЫЕ ПРЕДПОСЫЛКИ РЕГРЕССИОННОГО АНАЛИЗА. ТЕОРЕМА ГАУССА- МАРКОВА

Пусть для оценки параметров линейной функции регрессии использовалась выборка, содержащая ***n*** пар значений переменных (x_i; y_i), где $i = 1, 2, \ldots n$.

Линейная регрессионная модель с двумя переменными имеет вид:

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad (1)$$

$$(i = 1, 2, \ldots, n)$$

Чтобы регрессионный анализ, основанный на МНК, давал наилучшие из всех возможных результаты, должны выполняться определенные условия относительно возмущения $\boldsymbol{\varepsilon}$ – ***основные предпосылки регрессионного анализа.***

Основные предпосылки регрессионного анализа:
(условия Гаусса- Маркова)

1. *Математическое ожидание возмущения* $\boldsymbol{\varepsilon_i}$ *в любом наблюдении должно быть равно нулю, т.е.*

$$M(\varepsilon_i)=0,$$
$$i = 1, 2, \ldots, n$$

2. *Дисперсия возмущения* ε_i *должна быть постоянной для всех наблюдений:*

$$D(\varepsilon_i) = D(\varepsilon_j) = \sigma^2$$

3. *Возмущения* ε_i *и* ε_j *должны быть некоррелированы между собой.*

4. *В модели (1) возмущение* ε_i *есть величина случайная,*

а объясняющая переменная x_i *- величина неслучайная.*

При выполнении основных предпосылок регрессионного анализа модель (1) называется **классической нормальной линейной регрессионной моделью.**

Наряду с условиями 1 – 4 обычно предполагается, что возмущение ε_i имеет *нормальное распределение.*

Оценкой модели (1) по выборке является уравнение регрессии:

$$\hat{y} = a + bx \qquad (2)$$

Параметры этого уравнения определяются по МНК.

Выборочной оценкой возмущения ε_i является отклонение e_i значения y_i переменной Y от значения $\hat{y}_i$, найденного по уравнению регрессии (2), т. е.

$$e_i = y_i - \hat{y}_i$$

? Являются ли оценки a и b параметров α и β «наилучшими»?

ТЕОРЕМА ГАУССА- МАРКОВА:

Если основные предпосылки регрессионного анализа (условия 1 - 4) выполняются, то оценки (a, b), сделанные с помощью МНК, являются наилучшими линейными несмещенными оценками, т.е. обладают свойствами несмещенности, эффективности и состоятельности.

1) **несмещенность**: $M(a)=\alpha$; $M(b)=\beta$

2) **эффективность**: имеют наименьшую дисперсию в классе всех линейных несмещенных оценок.

3) **состоятельность**: при достаточно большом n оценки a и b близки к (α, β);

(при увеличении объема выборки надежность оценок увеличивается).

2. ГЕТЕРОСКЕДАСТИЧНОСТЬ. ТЕСТ ГОЛДФЕЛДА–КВАНДТА

Одна из предпосылок регрессионного анализа:

Дисперсия возмущения ε_i должна быть постоянна для всех наблюдений.

Выполнимость данной предпосылки называется **гомоскедастичностью**.

Гетероскедастичность – это нарушение основной предпосылки регрессионного анализа о постоянстве дисперсий случайных возмущений.

Основные причины появления гетероскедастичности

- **Ошибки спецификации модели** (неправильный отбор объясняющих переменных или выбор формы уравнения регрессии).
- **Нарушение принципа однородности выборки** (большие различия между наименьшими и наибольшими значениями наблюдений выборки).

Последствия гетероскедастичности

Выводы, получаемые на основе ***t*** - и ***F***- статистик, а также интервальные оценки будут ненадежными.

Для обнаружения гетероскедастичности используют:

- ***графический анализ***
- ***статистические тесты***.

I. *Графический анализ возмущений*

Для проведения графического анализа по оси абсцисс откладываются значения (x_i) объясняющей переменной X, а по оси ординат либо отклонения e_i, либо квадраты e_i^2, i =1,2,..n.

$e_i = y_i - \hat{y}_i$ – *оценки возмущений* ε_i, *полученные из выборочного уравнения регрессии.*

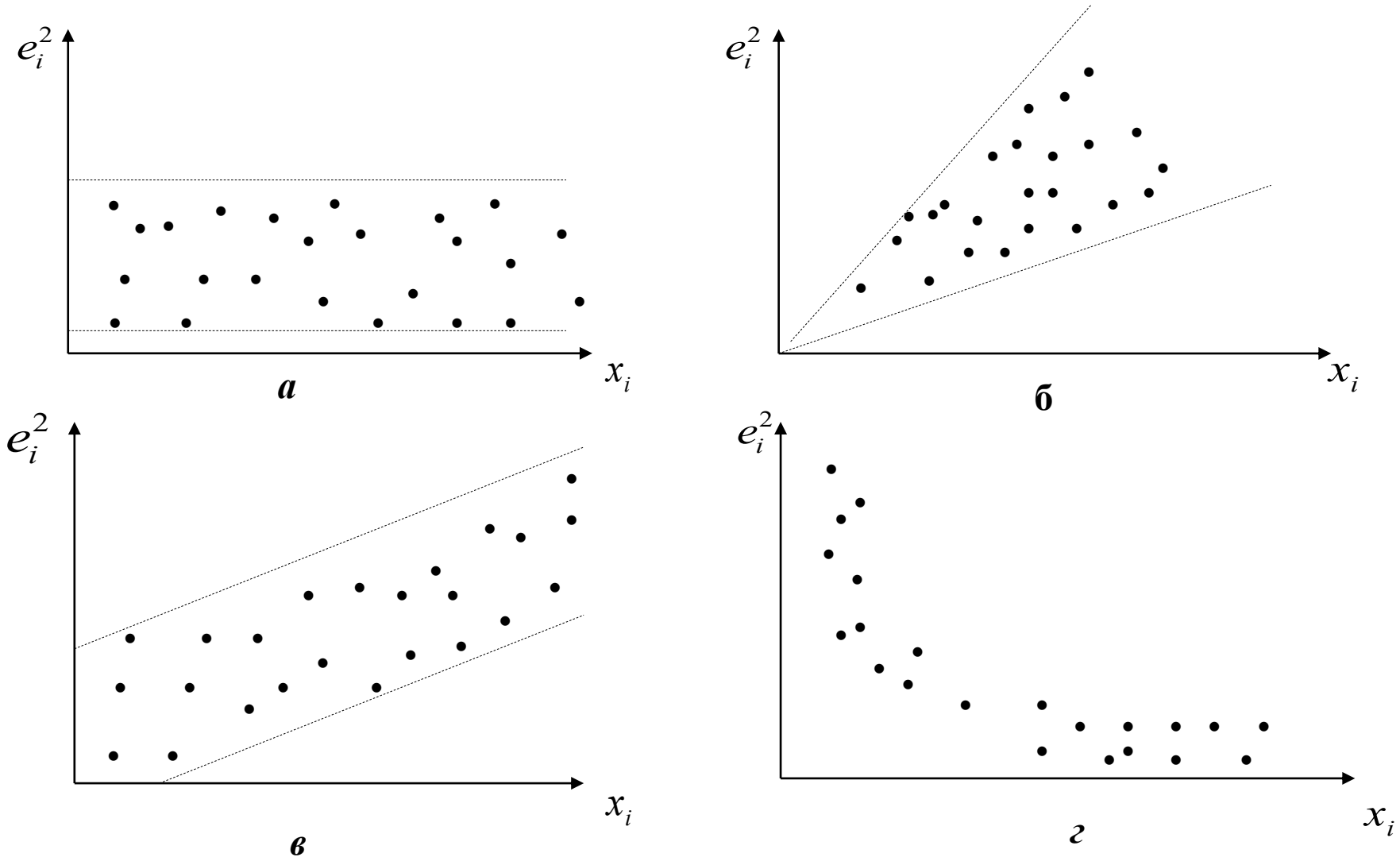

Рис. 1. Графики зависимости квадратов возмущений от значений X

На ***рис.а*** «облако» случайных отклонений $\boldsymbol{e_i^2}$ находится внутри полосы постоянной ширины, параллельной оси абсцисс. Это говорит о независимости дисперсии от значений переменной *X* и их постоянстве, т. е. выполняется условие *гомоскедастичности.*

На ***рис.б*** дисперсия $\boldsymbol{e_i^2}$ растет по мере увеличения значений переменой ***X***.

На ***рис.в*** отражена линейная, ***г*** – гиперболическая зависимости между квадратами отклонений и значениями переменной ***X***.

Ситуации на ***рис. б-г*** отражают большую вероятность наличия г*етероскедастичности*.

II. Статистические тесты на гетероскедастичность

- тест ранговой корреляции Спирмена;
- тест Голдфелда-Квандта;
- тест Глейзера;
- тест Уайта.

Все тесты используют в качестве нулевой гипотезы H_0 ***гипотезу об отсутствии гетероскедастичности.***

Тест Голдфелда - Квандта

При проведении проверки по этому тесту предполагается:

1) среднее квадратическое отклонение σ_i возмущения ε_i пропорционально значению x_i объясняющей переменной X в этом наблюдении.

2) случайное возмущение ε_i имеет нормальное распределение.

Тест включает шаги:

1. Все n наблюдений упорядочиваются по возрастанию переменной X.

2. Вся упорядоченная выборка после этого разбивается на три подвыборки размерностей m, $(n - 2m)$, m соответственно.

3. Оцениваются отдельные регрессии для первой подвыборки (m первых наблюдений) и для третьей подвыборки (m последних наблюдений). Средние $(n - 2m)$ наблюдений отбрасываются.

4. Строится F-статистика:

$$F = \frac{S_3/(m-p-1)}{S_1/(m-p-1)} = \frac{S_3}{S_1},$$

где $S_1 = \sum_{i=1}^{m} e_i^2$ – сумма квадратов

отклонений для первой выборки;

$S_3 = \sum_{i=n-m+1}^{n} e_i^2$ – сумма квадратов

отклонений для 3-ей выборки.

p – количество объясняющих переменных в уравнении регрессии;

(***m*** **–** ***p*** **–** **1**) – число степеней свободы соответствующих сумм.

Построенная F-статистика имеет распределение Фишера с числом степеней свободы $k_1 = k_2 = m - p - 1$.

5. Если $\boldsymbol{F > F_{табл.}}$, то гипотеза H_0
об отсутствии гетероскедастичности **отклоняется**.

3. АВТОКОРРЕЛЯЦИЯ. КРИТЕРИЙ ДАРБИНА – УОТСОНА

Основная предпосылка регрессионного анализа:

Возмущения ε_i и ε_j должны быть некоррелированы между собой

Если данное условие выполняется, то говорят об отсутствии ***автокорреляции*.**

Автокорреляция (последовательная корреляция) - это корреляция между наблюдаемыми показателями, упорядоченными во времени или в пространстве.

Последствия автокорреляции

Оценки параметров, полученные по МНК, перестают быть эффективными, а их стандартные ошибки рассчитываются некорректно (занижаются).

Вследствие этого ухудшаются прогнозные качества модели.

Методы определения автокорреляции:

- графический метод;
- критерий Дарбина-Уотсона.

I. Графический метод основан на построении последовательно-временных графиков.

По оси абсцисс откладываются либо время (момент) получения статистических данных, либо порядковый номер наблюдения,

а *по оси ординат* – оценки возмущений e_i.

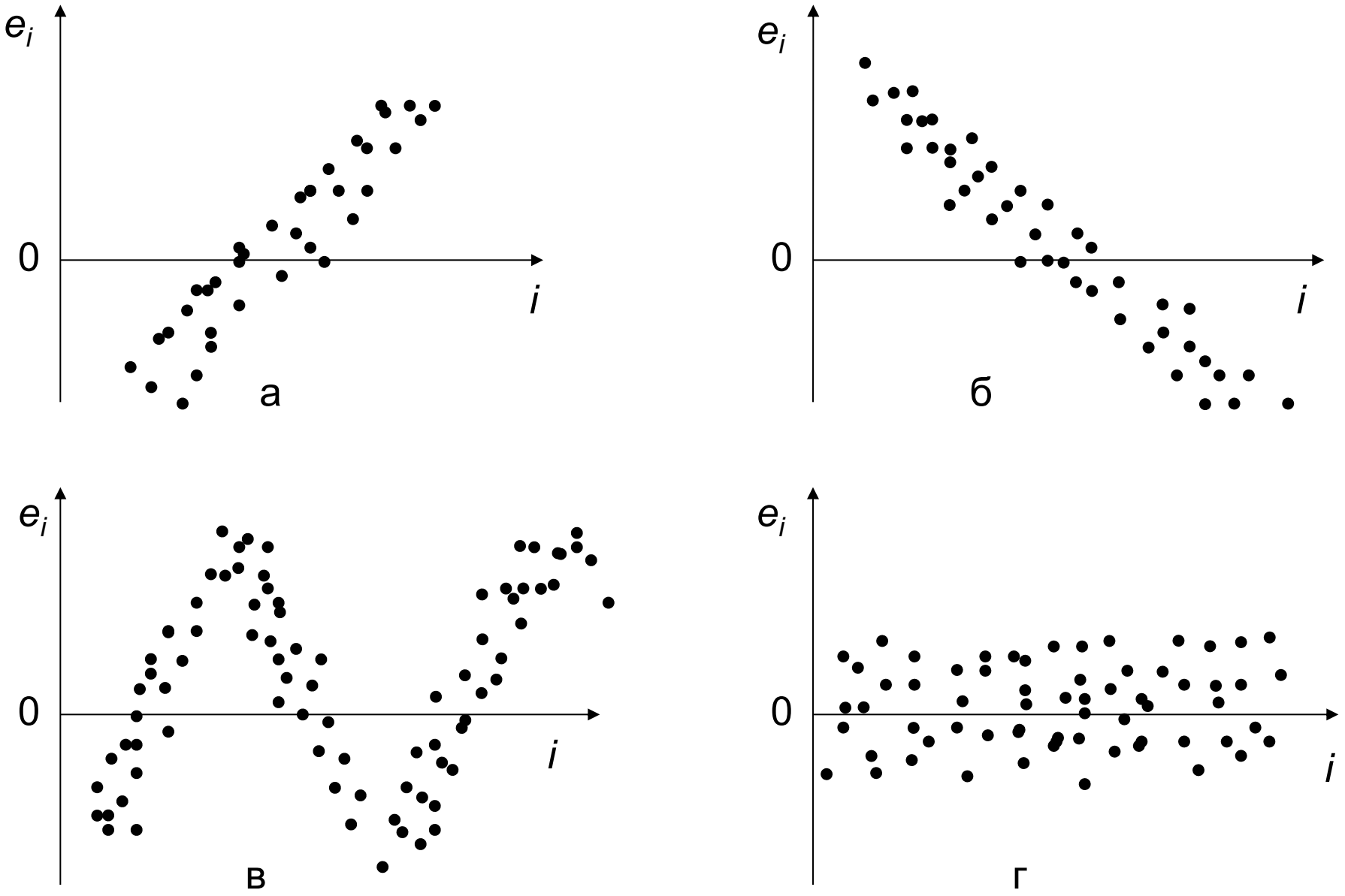

Рис.3. Графики зависимости отклонений от номера наблюдений

На рис. *а - в* имеются определенные связи между отклонениями:

а – нарастающая тенденция в остатках;

б – убывающая тенденция в остатках;

в – циклические колебания в остатках,

т.е. *автокорреляция* имеет место.

Отсутствие зависимости на рис. *г* скорее всего свидетельствует об отсутствии автокорреляции.

II. Критерий Дарбина-Уотсона определяет наличие автокорреляции

между соседними отклонениями, т.е. e_i и e_{i-1}.

В тесте используется статистика вида:

$$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

Статистика d тесно связана с коэффициентом корреляции между соседними наблюдениями:

$$d \approx 2\left(1 - r_{e_i e_{i-1}}\right),$$

где $r_{e_i e_{i-1}}$ – выборочный коэффициент корреляции между e_i и e_{i-1}

$$\textit{Если } r_{e_i e_{i-1}} \approx 0 \ (\textit{автокорреляция отсутствует}),$$

$$\textit{то } d \approx 2.$$

$$\textit{Если } r_{e_i e_{i-1}} \approx 1 \ (\textit{положительная автокорреляция}),$$

$$\textit{то } d \approx 0.$$

$$\textit{Если } r_{e_i e_{i-1}} \approx -1 \ (\textit{отрицательная автокорреляция}),$$

$$\textit{то } d \approx 4.$$

$$\textit{Таким образом, } 0 \le d \le 4.$$

Для более точного определения отсутствия или наличия автокорреляции была построена таблица критических точек распределения Дарбина-Уотсона.

По таблице для заданного уровня значимости $\boldsymbol{\alpha}$, числа наблюдений $\boldsymbol{n}$ и количества объясняющих переменных $\boldsymbol{p}$ определяются два пороговых значения:

$\boldsymbol{d_H}$ - *нижняя граница* и $\boldsymbol{d_B}$ - *верхняя граница*.

По этим значениям отрезок **[0; 4]** разбивается на пять зон. В зависимости от того, в какую зону попадает расчетное значение критерия, осуществляют выводы по правилу:

Если

1) $\boldsymbol{0 \leq d < d_н}$ - *существует положительная автокорреляция;*

2) $\boldsymbol{d_н \leq d < d_в}$ -

вывод о наличии автокорреляции не определен;

3) $\boldsymbol{d_в \leq d < 4-d_в}$ - *автокорреляция отсутствует;*

4) $\boldsymbol{4-d_в \leq d < 4-d_н}$ -

вывод о наличии автокорреляции не определен;

5) $\boldsymbol{4-d_н \leq d \leq 4}$ -

существует отрицательная автокорреляция.

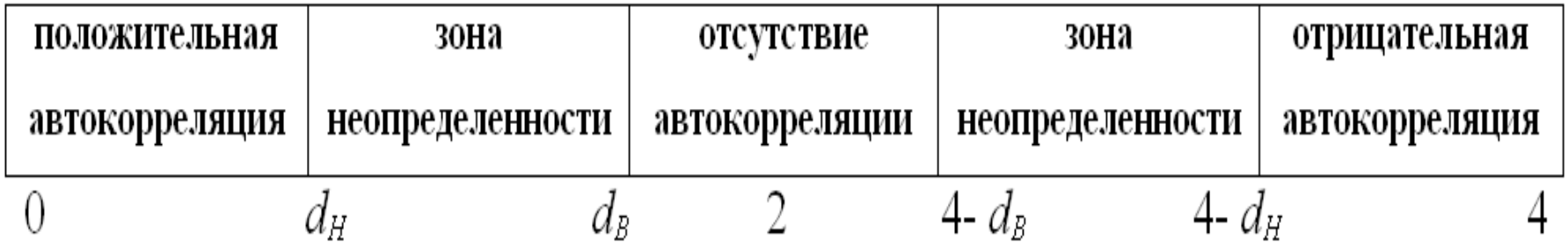

Рис.3 - Графическое изображение результата Дарбина -Уотсона