Проверить модели построенные в лабораторных 5 и 6 на

Автокорреляцию

Гетероскедастичность

Мультиколлинеарность

<u>Автокорреляция: понятие и последствия, вопросы выявления, методы коррекции автокорреляции.</u>

Ключевые понятия: автокорреляция; графический метод, статистика Дарбина-Уотсона DW, метод рядов, автокорреляционные функции, тест Бреуша-Годфри BG(k);

Важной предпосылкой построения качественной регрессионной модели по МНК является независимость значений случайных отклонений ε_i от значений отклонений во всех других наблюдениях, что гарантирует отсутствие коррелированности между любыми отклонениями и, в частности, между соседними отклонениями.

Автокорреляция (последовательная корреляция) определяется как корреляция между наблюдаемыми показателями, упорядоченными во времени (временные ряды) или в пространстве (перекрестные ряды). Понятие автокорреляции остатков (случайных отклонений) встречается в регрессионном анализе при построении моделей на основе временных рядов и очень редко при использовании перекрестных данных для проверки возможных ошибок спецификации (однако, в общем случае, такая проверка не имеет смысла для моделей пространственных данных).

Различают положительную и отрицательную автокорреляцию. Для случайных отклонений в регрессионных моделях для экономических показателей, как правило, характерна положительная автокорреляция, нежели отрицательная автокорреляция. В большинстве случаев положительная автокорреляция вызывается направленным постоянным воздействием некоторых неучтенных в регрессии факторов. Например, Y спрос на прохладительные напитки; X ежемесячный располагаемый доход. Фактические точки наблюдений и трендовая линейная модель представлены на рисунке.

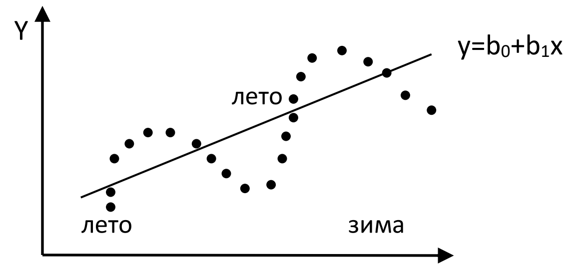


Рис.3.

Точки наблюдений в этом случае обычно будут превышать трендовую линию в летние периоды и будут ниже ее в зимние (что видно и из графика).

Отрицательная автокорреляция означает, что за положительным отклонением следует отрицательное, и наоборот. Такая ситуация может иметь место, если в рамках того же примера зависимость спроса на прохладительные напитки *Y* от доходов *X* рассматривать по сезонным данным (зима-лето). Вариант схемы рассеивания точек при отрицательной автокорреляции может выглядеть следующим образом (Рис.4).

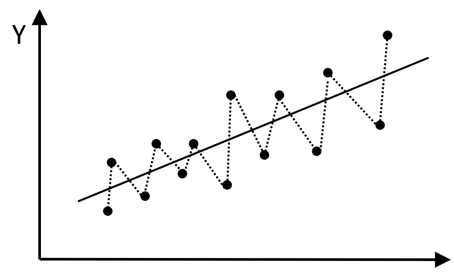


Рис. 4.

Причины автокорреляции:

1. Ошибки спецификации. Пропущена в модели какая-либо важная объясняющая переменная либо неправильный выбор формы зависимости

обычно приводят к системным отклонениям точек наблюдения от линии регрессии, что может обусловить автокорреляцию.

2. Инерция. Многие экономические показатели (инфляция, безработица, ВНП и т.д.) обладают определенной цикличностью, связанной с волнообразностью деловой активности. Поэтому изменение показателей обладает определенной инертностью.

3. Эффект паутины. Во многих производственных и других сферах экономические показатели реагируют на изменение экономических условий с запаздыванием (временным лагом).

4. Сглаживание данных. Зачастую данные по некоторому продолжительному временному периоду получают усреднением данных по составляющим его интервалам. Это может привести к определенному сглаживанию колебаний, которые имелись внутри рассматриваемого периода, что в свою очередь может служить причиной автокорреляции.

Последствия автокорреляции:

1. Оценки параметров модели остаются линейными и несмещенными, но перестают быть эффективными, то есть перестают быть *BLUE*-оценками.
2. Оценка дисперсии случайных отклонений является смещенной , чаще заниженной , поэтому R^2 является зависимой оценкой и завышен.
3. Дисперсии оценок являются смещенными, и, как правило, их значения также занижены, что приводит к росту t-статистики и переоценке статистической значимости параметров модели. Выводы по t- и F-статистикам ненадежны, ухудшается прогнозное качество модели.

Обнаружение автокорреляции:

Существует несколько методов, позволяющих обнаружить автокорреляцию.

1. Графический метод. Есть ряд вариантов графического определения автокорреляции. Один из них увязывает отклонения e_i с моментами их получения $i = 1,2\ldots,n$. При этом по оси абсцисс откладывают либо время получения статистических данных, либо порядковый номер наблюдения, а по оси ординат – отклонения ε_i, либо оценки отклонений e_i .

Естественно предположить, что на рисунках 6а-г имеются определенные связи между отклонениями, т.е. автокорреляция имеет место. Отсутствие зависимости на рисунке 6д скорее свидетельствует об отсутствии автокорреляции.

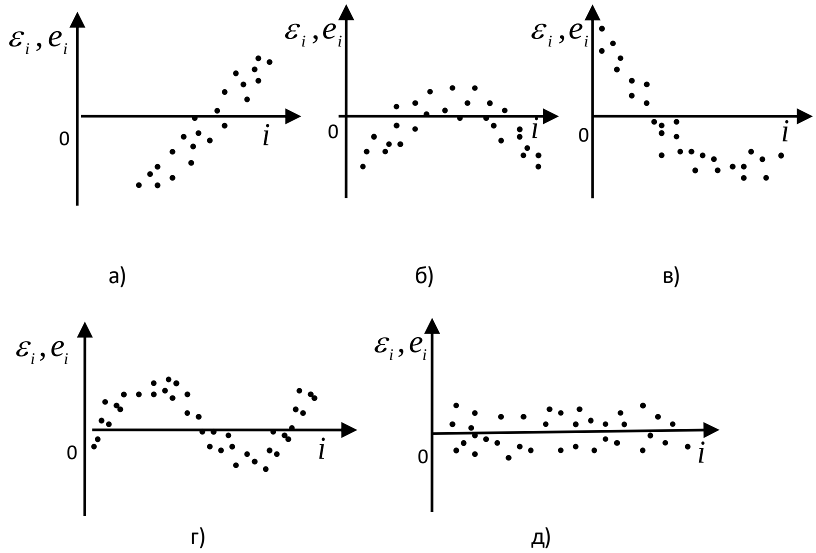


Рис. 6.

Для случая 6б отклонения сначала являются отрицательными, затем положительными, затем снова отрицательными. Это свидетельствует о наличии между отклонениями определенной зависимости, более того, можно утверждать, что в этом случае имеет место положительная автокорреляция. Она становится более наглядной, если построить график зависимости e_i от e_{i-1} (рис. 7).

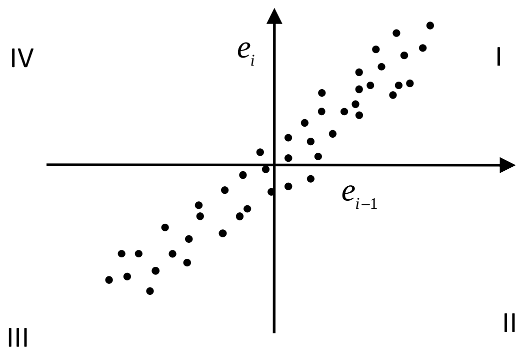


Рис. 7.

Большинство точек на этом графике расположено в I и III четвертях декартовой системы координат, подтверждая положительную зависимость между соседними отклонениями.

2. Метод рядов или метод Сведа-Эйзенхарта. Этот метод достаточно прост: последовательно определяются знаки отклонений e_t, t=1,2…T. Например,

(-----)(+++++++)(---)(++++)(-),

Т.е. 5 «-», 7 «+», 3 «-», 4 «+», 1 «-» при 20 наблюдениях.

Ряд определяется как непрерывная последовательность одинаковых знаков. Количество знаков в ряду называется длиной ряда. Визуальное распределение знаков свидетельствует о неслучайном характере связей между отклонениями. Если рядов слишком мало по сравнению с количеством наблюдений n, то вполне вероятна положительная автокорреляция. Если же рядов слишком много, то вероятна отрицательная автокорреляция.

3. Критерий Дарбина-Уотсона

Наиболее известным критерием обнаружения автокорреляции первого порядка является статистика (критерий) Дарбина- Уотсона (*DW*):

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

Между критерием Дарбина–Уотсона и коэффициентом автокорреляции остатков первого порядка имеет место следующее соотношение:

$$DW \approx 2 \cdot (1 - corr(e_t; e_{t-1}))$$

При проверке отсутствия автокорреляции случайных отклонений моделей с помощью статистики Дарбина-Уотсона согласно общей схемы проверки гипотез формулируют:

H_0: отсутствие автокорреляции остатков

H_1: есть автокорреляция остатков

Учитывая свойства коэффициента корреляции, получаем, что значение критерия Дарбина-Уотсона изменяется в пределах $0 \le DW \le 4$; при $corr(e_t; e_{t-1}) = 0$, т.е. $DW = 2$ автокорреляция отсутствует; при $corr(e_t; e_{t-1}) = 1$, т.е. $DW = 0$ присутствует положительная автокорреляция; при $corr(e_t; e_{t-1}) = -1$, т.е. $DW = 4$ присутствует отрицательная автокорреляция. Статистика *DW* характеризуется наличием, так называемых, зон неопределенности или «мертвых зон», разделяющих область принятия гипотезы H_0 и критические области. При попадании значения статистики

Дарбина-Уотсона в такую зону, вывод на основе *DW* не определен и необходимо использовать другие методы выявления автокорреляции. Алгоритм выявления автокорреляции остатков на основе критерия Дарбина–Уотсона следующий: после вычисления значения статистики *DW,* по таблице критических точек для распределения Дарбина–Уотсона для заданного числа наблюдений *n*, числа независимых переменных модели *m* и уровня значимости α находят значения двух границ интервалов, определяющих «мертвые зоны»: d_L (*low*) и d_U (*upper*). По этим значениям числовой промежуток [0;4] разбивают на пять отрезков. Принятие или отклонение каждой из гипотез с вероятностью (1–a) рассматривается на рис. 8.

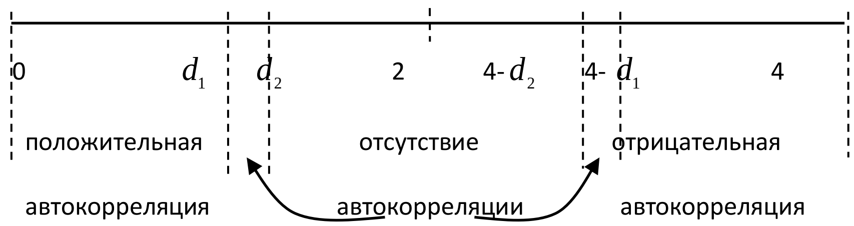


Рис. 8. Механизм проверки гипотезы о наличии автокорреляции остатков

На практике в некоторых случаях, если фактическое значение критерия *DW* попадает в зону неопределенности, то предполагают существование автокорреляции остатков, т.е. отклоняют гипотезу H_0. Таким образом, статистика Дарбина-Уотсона *DW* имеет определенные ограничения в применении, которые можно определить как ее недостатки:

(a) Статистика имеет зоны неопределенности, при попадании в которые в общем случае сделать вывод не представляется возможным;

(b) Статистика неприменима к моделям, включающим в качестве независимых переменных лаговые значения результативного признака, т. е. к моделям авторегрессии;

(c) Методика расчета и использования критерия Дарбина-Уотсона направлена только на выявление автокорреляции первого порядка. При проверке остатков на автокорреляцию более высоких порядков следует применять другие методы;

(d) Статистика Дарбина-Уотсона дает достоверные результаты только для больших выборок.

4. Тест серий (тест Бреуша-Годфри). Используется для больших выборок и выявления автокорреляции высоких порядков. Тест основан на следующей идее: если имеется корреляция между соседними наблюдениями, то естественно ожидать, что в уравнении:

$$e_t = \rho_1 \cdot e_{t-1} + \rho_2 \cdot e_{t-2} + \ldots + \rho_k \cdot e_{t-k} + v_t\,,\; t = \overline{1,n}$$

где e_t – случайные отклонения исходной модели регрессии, которая тестируется на автокорреляцию, коэффициент ρ_k окажется значимо отличающимся от нуля. Таким образом, гипотеза формулируется:

H_0: $\rho_1=\rho_2=\ldots=\rho_k=0$ (отсутствует автокорреляция)

H_1: $\rho_k \neq 0$ (присутствует автокорреляция порядка k)

Практическое применение теста для проверки гипотезы заключается в оценивании методом наименьших квадратов вспомогательной регрессии, а общая схема теста выглядит следующим образом:

1. Оценка исходной регрессии и выделение ряда случайных отклонений e_t;
2. Оценка вспомогательной регрессии e_t на все экзогенные факторы исходной модели, а также лаги отклонений e_t по проверяемый порядок k включительно:

$$e_t = \alpha_0 + \alpha_1 \cdot x_{1t} + \ldots + \alpha_m \cdot x_{mt} + \rho_1 \cdot e_{t-1} + \rho_2 \cdot e_{t-2} + \ldots + \rho_k \cdot e_{t-k} + v_t$$

3. Расчет статистики теста $BG(k)$ на основе коэффициента детерминации R^2 вспомогательной модели:

$$BG(k) = (n-k) \cdot R^2 \sim \chi^2_{\alpha;k}$$

Нулевая гипотеза об отсутствии автокорреляции k-го порядка принимается при условии, что $BG(k) < \chi^2_{\alpha;k}$.

Для проверки гипотезы можно использовать так же F-статистику, сравнивающую между собой коэффициенты детерминации вспомогательной модели теста и этой же модели без лагов случайных отклонений: $e_t = \alpha_0 + \alpha_1 \cdot x_{1t} + \ldots + \alpha_m \cdot x_{mt} + u_t$. При выполнении предпосылки МНК об отсутствии линейной зависимости между случайными отклонениями и экзогенными переменными, коэффициент детерминации последней модели сравним с нулем.

Пример

В таблице 2.1 приведены поквартальные данные об объеме валового продукта (переменная GDP – Gross Domestic Product less Net Exports, billions of dollars) и потребительских расходах продукта (переменная CN – Personal consumption expend, billions of dollars) в США в период с 1980 по 1990 год. Оцените линейную регрессионную зависимость расходов от объема валового продукта, проверьте гипотезу о некоррелированности случайных отклонений полученной модели.

Таблица 2.1

Годы	Квартал	GDP	CN	Годы	Квартал	GDP	CN
1980	I	4578,95	2952,51	**1985**	III	5444,98	3537,53
	II	4454,66	2885,44		IV	5497,56	3549,72
	III	4425,76	2914,76	**1986**	I	5535,67	3580,59
	IV	4523,03	2949,76		II	5575,37	3619,53
1981	I	4618,81	2962,31		III	5631,77	3683,03
	II	4585,22	2964,51		IV	5651,56	3707,86
	III	4643,24	2978,25	**1987**	I	5687,69	3710,55
	IV	4595,09	2954,89		II	5744,75	3758,01
1982	I	4522,93	2972,57		III	5785,45	3799,67
	II	4534,90	2980,91		IV	5880,99	3807,25
	III	4542,65	3000,34	**1988**	I	5892,92	3873,86
	IV	4546,18	3050,73		II	5941,45	3900,25
1983	I	4600,89	3078,67		III	5975,76	3932,11
	II	4733,10	3142,81		IV	6054,45	3977,80
	III	4839,56	3192,84	**1989**	I	6101,13	3992,15
	IV	4950,04	3245,52		II	6125,10	4008,37
1984	I	5081,19	3279,88		III	6159,00	4043,78
	II	5175,63	3324,41		IV	6173,56	4058,42
	III	5224,01	3347,50	**1990**	I	6237,52	4091,92

	IV	5269,61	3390,93		II	6251,30	4103,67
1985	I	5302,11	3442,15		III	6245,69	4119,08
	II	5366,97	3473,94		IV	6162,51	4084,33

Источник данных: Bureau of Economic Analysis (www.bea.gov), U.S. national income & product accounts

Используя принятые в таблице обозначения, оценим параметры регрессии $\hat{CN}_t = \beta_0 + \beta_1 \cdot GDP_t$:

Модель 1

Зависимая (эндогенная) переменная *CN*

Количество наблюдений 44

Переменная	Коэффициент	Ст. ошибка	t-стат.	P-вероятн.
Константа C	-54,24070	46,86539	-1,157372	0,2537
Независимая (экзогенная) переменная *GDP*	0,663393	0,008719	76,08230	0,0000
Коэффициент детерминации	0,992797	F-статистика		5788,516

Полученные результаты указывают на статистическую значимость коэффициента при экзогенной переменной; в тоже время, исходя из чрезмерно высоких значений коэффициента детерминации и *t*-статистики коэффициента при экзогенной переменной, по формальным признакам можно сделать предварительный вывод о наличии коррелированности случайных отклонений модели. В этом случае, полученные значения дисперсий смещены (стандартная ошибка регрессии, стандартные ошибки коэффициентов) и нами могут быть сделаны ложные выводы о качестве построенной модели.

Статистика Дарбина-Уотсона (Durbin-Watson). Найдем значение статистики DW, для того, чтобы проверить гипотезу об отсутствии коррелированности случайных отклонений модели (автокорреляции первого порядка):

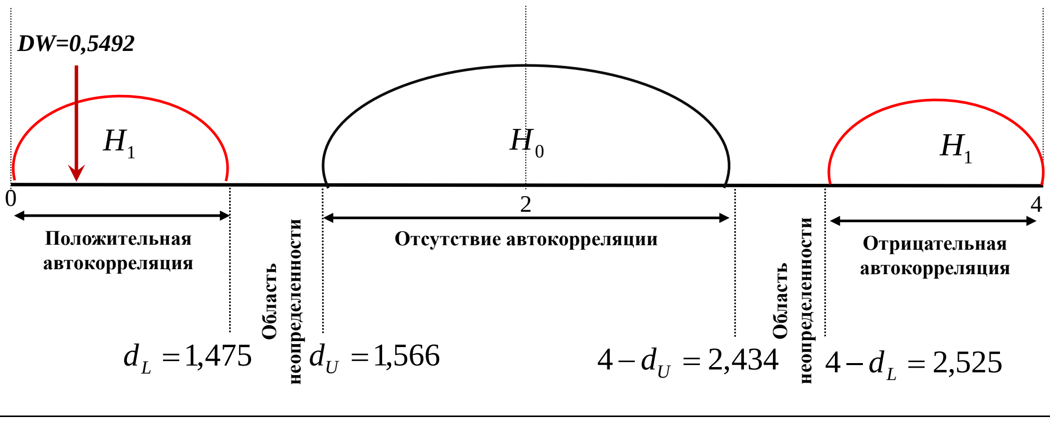
Вычисляем соответствующие значения сумм:

$$\sum_{t=1}^{44} e_t^{\ 2} = 55476{,}5627 \quad ; \sum_{t=2}^{44}\left(e_t - e_{t-1}\right)^2 = 30468{,}4316$$

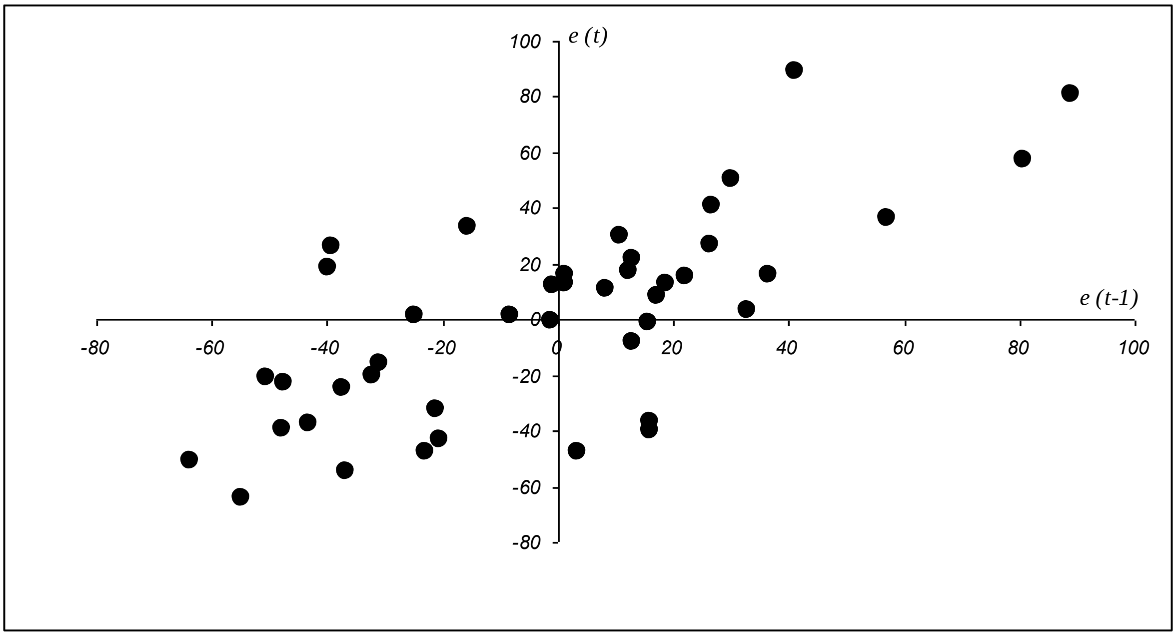
Находим значение статистики Дарбина-Уотсона (Durbin-Watson):

$$DW = \frac{\sum_{t=2}^{44}\left(e_t - e_{t-1}\right)^2}{\sum_{t=1}^{44} e_t^{\ 2}} = \frac{30468{,}4316}{55476{,}5627} = 0{,}549213$$

С помощью таблицы критических точек в таблице распределения DW для уровня значимости $\alpha = 0{,}05$ и значений объема выборки $n = 44$, количества объясняющих переменных в модели $m = 1$ определяем $d_L = 1{,}475; d_U = 1{,}566$ и делаем вывод о наличии положительной автокорреляции случайных отклонений.



Графический метод. Визуальный анализ корреляционного поля случайных отклонений e_t и их лаговых значений e_{t-1} так же указывает на закономерность в распределении точек $\left(e_{t-1}; e_t\right)$ и наличие между случайными отклонениями положительной линейно зависимости (автокорреляции). Коэффициент парной корреляции между ними равен $corr(e_{t-1}; e_t) = 0,71776$ и является статистически значимым.



Тест Бреуша-Годфри (Breusch-Godfrey). Для проверки гипотезы об отсутствии автокорреляции первого порядка случайных отклонений исходной модели построим вспомогательную модель вида $\hat{e}_t = \beta_0 + \beta_1 \cdot GDP_t + \delta_1 \cdot e_{t-1}$:

LM-тест Бреуша-Годфри (Breusch-Godfrey), лаг k=1

H_0 : случайные отклонения модели не коррелированы

H_1 : присутствует автокорреляция первого порядка случайных отклонений модели

F-статистика теста	42,5352	$F_{0,05}(1,40)$ =4,085	P-вероятность	0,0000
Статистика $(n-1)\cdot R^2$	22,1604	$\chi^2{}_{0,05}(1) = 3,84$	P-вероятность	0,0000

Вспомогательное регрессионное уравнение

Зависимая (эндогенная) переменная e_t

Количество наблюдений 43

Переменная	Коэффициент	Ст. ошибка	t-стат.	P-вероятн.
C	5,6300	33,8229	0,1665	0,8686
GDP	-0,0008	0,0063	-0,1208	0,9045
e_{t-1}	0,7285	0,1118	6,5179	0,0000
Коэффициент детерминации	0,515358	F-статистика		21,2676

Для проверки гипотезы о некоррелированности случайных отклонений модели в тесте Бреуша-Годфри используем значение статистики $BG(1)=(n-1)\cdot R^2$, где n – число наблюдений, которое было использовано при построении исходной модели; R^2 – коэффициент детерминации вспомогательной модели. С помощью таблицы критических значений χ^2-распределения находим критическую точку $\chi^2{}_{0,05}(1)=3,84$. Поскольку значение статистики $(n-1)\cdot R^2 = 22,1604 > 3,84 = \chi^2{}_{0,05}(1)$, то H_0 отклоняется при этом уровне значимости. Следовательно, предпосылка Гаусса-Маркова о некоррелированности случайных отклонений модели не выполняется, в модели присутствует автокорреляция первого порядка случайных отклонений. P-вероятность для статистики $(n-1)\cdot R^2$также показывает, что гипотеза H_0 отклоняется.

Для проверки гипотезы можно использовать так же F-статистику, сравнивающую между собой коэффициенты детерминации вспомогательной модели и модели $\hat{e}_t = \beta_0 + \beta_1 GDP_t$, коэффициент детерминации которой предполагается равным нулю:

$$F_{набл} = \frac{R_1{}^2 - R_2{}^2}{1-R_1{}^2}\cdot\frac{n'-m-1}{k} = \frac{0,515358-0}{1-0,515358}\cdot\frac{43-2-1}{1} = 42,5352$$

Находим значение критической точки в таблице распределения Фишера для уровня значимости $\alpha = 0,05$ и значений степеней свободы

$\nu_1 = k = 1,\ \nu_2 = n' - m - 1 = 40$: $F_{крит} = 4{,}085$ (объем выборки при проверке гипотезы определяется по количеству наблюдений для вспомогательной модели, т.е. $n' = n - l = n - 1 = 43$). Поскольку $F_{набл} > F_{крит}$, то есть основания для отклонения нулевой (основной) гипотезы, согласно которой $R_1^2 = R_2^2$, т.е. введение лага e_{t-1} повышает качество вспомогательной модели, т.е. присутствует автокорреляция первого порядка случайных отклонений исходной модели.

Аналогичным образом, построив вспомогательную модель вида $\hat{e}_t = \beta_0 + \beta_1 \cdot GDP_t + \delta_1 \cdot e_{t-1} + \delta_2 \cdot e_{t-2}$, отклоняем гипотезу об отсутствии автокорреляции второго порядка случайных отклонений исходной модели:

LM-тест Бреуша-Годфри (Breusch-Godfrey), лаг k=2

H_0 : случайные отклонения модели не коррелированы

H_1 : присутствует автокорреляция второго порядка случайных отклонений модели

F-статистика теста	21,0009	$F_{0,05}(2{,}39)$ =3,238	Р-вероятность	0,0000
Статистика $(n-1)\cdot R^2$	21,7782	$\chi^2_{0,05}(2) = 5{,}99$	Р-вероятность	0,0000

Вспомогательное регрессионное уравнение

Зависимая (эндогенная) переменная e_t

Количество наблюдений 42

Переменная	Коэффициент	Ст. ошибка	t-стат.	Р-вероятн.
C	3,2851	35,5745	0,0923	0,9269
GDP	-0,0004	0,0066	-0,0567	0,9550
e_{t-1}	0,8056	0,1638	4,9172	0,0000

e_{t-2}	-0,1042	0,1638	-0,6361	0,5285

Коэффициент детерминации	0,518529	F-статистика	13,6416

Для проверки гипотезы об отсутствии автокорреляции второго порядка находим значение статистики $BG(2) = (n-2) \cdot R^2$. С помощью таблицы критических значений χ^2-распределения находим критическую точку $\chi^2_{0,05}(2) = 5{,}99$. Поскольку значение статистики $BG(2) = (n-2) \cdot R^2 = 21{,}7782 > 5{,}99 = \chi^2_{0,05}(2)$, то H_0 отклоняется и в исходной модели присутствует автокорреляция второго порядка случайных отклонений.

Используя для проверки гипотезы F-статистику, получаем $F_{набл} = 21{,}0009$. Находим значение критической точки в таблице распределения Фишера для уровня значимости $\alpha = 0{,}05$ и значений степеней свободы $\nu_1 = k = 2,\ \nu_2 = n - m - 1 = 39$: $F_{крит} = 3{,}238$. Поскольку $F_{набл} > F_{крит}$, то подтверждается наличие у исходной модели автокорреляция второго порядка случайных отклонений.

Метод рядов. Дополним наше исследование случайных отклонений модели (см. таблицу 2.2) на автокорреляцию методом рядов или методом Сведа-Эйзенхарта (Swed-Eisenhart):

Таблица 2.2

Год		e_t	Год		e_t	Год		e_t
Квартал			Квартал			Квартал		
1980	I	**-30,895**	**1984**	I	**-36,708**	**1988**	I	18,776
	II	**-15,512**		II	**-54,829**		II	12,971
	III	32,980		III	**-63,834**		III	22,070

	IV	3,452		IV	**-50,654**		IV	15,558
1981	I	**-47,538**	**1985**	I	**-20,995**	**1989**	I	**-1,059**
	II	**-23,054**		II	**-32,232**		II	**-0,741**
	III	**-47,805**		III	**-20,394**		III	12,180
	IV	**-39,222**		IV	**-43,085**		IV	17,161
1982	I	26,328	**1986**	I	**-37,497**	**1990**	I	8,231
	II	26,728		II	**-24,893**		II	10,839
	III	41,016		III	1,191		III	29,971
	IV	89,064		IV	12,893		IV	50,402
1983	I	80,710	**1987**	I	**-8,386**			
	II	57,143		II	1,221			
	III	36,548		III	15,881			
	IV	15,936		IV	**-39,920**			

Находим n_1 – количество случайных отклонений с положительным знаком, n_2 – с отрицательным знаком и k - количество рядов (последовательностей случайных отклонений одного знака, для удобства выделенных в таблице разным цветом): $n_1 = 24, n_2 = 20, k = 12$.

При объеме выборки *n<40* для определения критических значений количества рядов можно воспользоваться соответствующей таблицей критических значений для уровня значимости *α=0,05*. По значениям n_1 и n_2 определяют нижнюю k_1 и верхнюю k_2 границы области принятия нулевой гипотезы об отсутствии коррелированности случайных отклонения модели. В нашем случае воспользуемся следующими формулами для расчета нижней и верхней границ:

$$k_1 = \left[M(k) - u_{\frac{\alpha}{2}} \cdot \sqrt{D(k)} \right], k_2 = \left[M(k) + u_{\frac{\alpha}{2}} \cdot \sqrt{D(k)} \right],$$

где *M(k)* – соответствующее математическое ожидание, *D(k)* – дисперсия, $u_{\frac{\alpha}{2}}$ – критическое значение нормального стандартизированного распределения (используем функцию Лапласа, $u_{\frac{0,05}{2}} = 1,96$), [.] – операция взятия целой части числа.

Вычисляем значения числовых характеристик и находим k_1, k_2:

$$M(k) = \frac{2 \cdot n_1 \cdot n_2}{n_1 + n_2} + 1 = \frac{2 \cdot 24 \cdot 20}{24 + 20} + 1 = 22,8182$$

$$D(k) = \frac{2 \cdot n_1 \cdot n_2 \cdot (2 \cdot n_1 \cdot n_2 - n_1 - n_2)}{(n_1 + n_2)^2 \cdot (n_1 + n_2 - 1)} = \frac{2 \cdot 24 \cdot 20 \cdot (2 \cdot 24 \cdot 20 - 24 - 20)}{(24 + 20)^2 \cdot (24 + 20 - 1)} =$$
$$= 10,5631$$

$$k_1 = \left[22,8182 - 1,96 \cdot \sqrt{10,5631}\right] = [16,448] = 16;$$
$$k_2 = \left[22,8182 + 1,96 \cdot \sqrt{10,5631}\right] = [29,1884] = 29$$

Поскольку $k = 12 < k_1 = 16$, то нулевая гипотеза о некоррелированности случайных отклонений модели отвергается в пользу альтернативного утверждения о том, что в исходной модели присутствует положительная автокорреляция первого порядка.

Таким образом, обобщив полученные результаты верификации модели (1), можно сделать общий вывод о наличии автокорреляции случайных отклонений, как первого, так и более высоких порядков. Изменим исходную модель (1) с целью коррекции автокорреляции случайных отклонений, введя в рассмотрение в качестве экзогенной переменной лаг переменной *CN,* т.е. преобразуя исходную модель в модель авторегрессии.

Модель 2

Зависимая (эндогенная) переменная *CN*

Количество наблюдений 43

Переменная	Коэффициент	Ст. ошибка	t-стат.	P-вероятн.

Константа *C*	-11,39092	30,59176	-0,372352	0,7116
Независимая (экзогенная) переменная CN_{t-1}	0,683044	0,086131	7,930272	0,0000
Независимая (экзогенная) переменная *GDP*	0,212583	0,056944	3,733176	0,0006
Коэффициент детерминации	0,997142	F-статистика		6978,156

Тест Бреуша-Годфри (Breusch-Godfrey). Для проверки гипотезы об отсутствии автокорреляции первого порядка случайных отклонений исходной модели построим вспомогательную модель вида $\hat{e}_t = \beta_0 + \beta_1 \cdot CN_{t-1} + \beta_2 \cdot GDP_t + \delta_1 \cdot e_{t-1}$:

LM-тест Бреуша-Годфри (Breusch-Godfrey), лаг k=1

H_0 : случайные отклонения модели не коррелированы

H_1 : присутствует автокорреляция первого порядка случайных отклонений модели

F-статистика теста	3,4614	$F_{0,05}(1{,}38)$ =4,098	Р-вероятность	0,0706
Статистика $(n-1)\cdot R^2$	3,5063	$\chi^2_{0,05}(1) = 3{,}84$	Р-вероятность	0,0611

Вспомогательное регрессионное уравнение

Зависимая (эндогенная) переменная e_t

Количество наблюдений 42

Переменная	Коэффициент	Ст. ошибка	t-стат.	Р-вероятн.
C	30,0964	27,7646	1,08398	0,2852
CN_{t-1}	0,06269	0,07897	0,79382	0,4322
GDP	-0,04596	0,05246	-0,87608	0,3865
e_{t-1}	0,19302	0,14235	1,35594	0,1831

Коэффициент детерминации	0,083484	F-статистика	1,1537

Для проверки гипотезы о коррелированности случайных отклонений модели (2) используем значение статистики $BG(1) = (n-1) \cdot R^2$, где n - число наблюдений для модели (2), Фактически из объема исходной выборки вычитается два наблюдения, на которые сократилась наша выборка вследствие перехода к авторегрессии в модели (2) и спецификации вспомогательной регрессии теста Бреуша-Годфри, R^2 - коэффициент детерминации вспомогательной модели, С помощью таблицы критических значений χ^2-распределения находим критическую точку $\chi^2_{0,05}(1) = 3,84$, Поскольку значение статистики $BG(1) = 3,50633 < 3,84 = \chi^2_{0,05}(1)$, то H_0 не отклоняется при этом уровне значимости, Следовательно, предпосылка Гаусса-Маркова о гомоскедастичности случайных отклонений модели выполняется. Р-вероятность для статистики $BG(1)$ показывает, что гипотеза H_0 будет отклоняться при уровне значимости $\alpha = 0,10$, т,е, в этом случае можно будет считать случайные отклонения автокоррелированными,

Для проверки гипотезы можно использовать так же F-статистику, сравнивающую между собой коэффициенты детерминации вспомогательной модели и модели $\hat{e}_t = \beta_0 + \beta_1 GDP_t + \beta_2 CN_{t-1}$, коэффициент детерминации которой предполагается равным нулю:

$$F_{набл} = \frac{R_1^2 - R_2^2}{1 - R_1^2} \cdot \frac{n-m-1}{k} = \frac{0,083484 - 0}{1 - 0,083484} \cdot \frac{42-3-1}{1} = 3,46136$$

Находим значение критической точки в таблице распределения Фишера для уровня значимости $\alpha = 0,05$ и значений степеней свободы $\nu_1 = k = 1,\ \nu_2 = n - m - 1 = 38$: $F_{крит} = 4,098$. Поскольку $F_{набл} < F_{крит}$, то нет оснований для отклонения нулевой (основной) гипотезы, согласно которой $R_1^2 = R_2^2$, т,е, введение лага e_{t-1} не улучшает качество вспомогательной модели, т.е. автокорреляция первого порядка случайных отклонений исходной модели (2) отсутствует.

Аналогичным образом, построив вспомогательную модель вида $\hat{e}_t = \beta_0 + \beta_1 \cdot CN_{t-1} + \beta_2 \cdot GDP_t + \delta_1 \cdot e_{t-1} + \delta_2 \cdot e_{t-2}$, принимаем гипотезу об отсутствии автокорреляции второго порядка случайных отклонений исходной модели при $\alpha = 0{,}05$ и отклоняем ее при $\alpha = 0{,}10$:

LM-тест Бреуша-Годфри (Breusch-Godfrey), лаг k=2

H_0 : случайные отклонения модели не коррелированы

H_1 : присутствует автокорреляция второго порядка случайных отклонений модели

F-статистика теста	2,8743	$F_{0,05}(2{,}36)$=3,259	P-вероятность	0,0695
Статистика $(n-1)\cdot R^2$	5,6455	$\chi^2{}_{0,05}(2) = 5{,}99$	P-вероятность	0,0594

Вспомогательное регрессионное уравнение

Зависимая (эндогенная) переменная e_t

Количество наблюдений 41

Переменная	Коэффициент	Ст. ошибка	t-стат.	P-вероятн.
C	25,1195	29,1453	0,8619	0,3945
CN_{t-1}	0,0536	0,0811	0,6605	0,5131
GDP	-0,0393	0,0543	-0,7238	0,4739
e_{t-1}	0,2443	0,1689	1,4462	0,1568
e_{t-2}	0,1587	0,1453	1,0921	0,2820
Коэффициент детерминации	0,137694	F-статистика		1,4371

Для проверки гипотезы об отсутствии автокорреляции второго порядка находим значение статистики $BG(2) = (n-2)\cdot R^2$. С помощью таблицы критических значений χ^2-распределения находим критическую точку

$\chi^2{}_{0,05}(2) = 5{,}99$. Поскольку значение статистики $BG(2) = 5{,}6455 < 5{,}99 = \chi^2{}_{0,05}(2)$, то H_0 отклоняется и в исходной модели отсутствует автокорреляция второго порядка случайных отклонений. Р-вероятность для статистики $BG(2)$ показывает, что гипотеза H_0 будет отклоняться при уровне значимости $\alpha = 0{,}10$, т.е. в этом случае можно будет считать случайные отклонения автокоррелированными.

Убедиться в том, что в скорректированной модели отсутствует автокорреляция более высоких порядков можно самостоятельно, используя соответствующие коэффициенты детерминации вспомогательных моделей: для $BG(3) \Rightarrow R^2 = 0{,}1761$, для $BG(4) \Rightarrow R^2 = 0{,}2074$, для $BG(5) \Rightarrow R^2 = 0{,}3095$ и т.д.

Таким образом, преобразование модели (1), заключавшееся в введении в состав экзогенных факторов авторегрессионной переменной, позволило скорректировать автокорреляцию первого и второго порядка при принятом уровне значимости $\alpha = 0{,}05$.

<u>Гетероскедастичность: понятие и последствия, вопросы выявления, методы коррекции гетероскедастичности.</u>

Ключевые понятия: гомоскедастичность, гетероскедастичность; графический метод, тесты Спирмена, Парка, Глейзера, Голдфелда-Квандта, Вайта, Бреуша-Пагана; ошибки спецификации; метод взвешенных наименьших квадратов (обобщенный метод наименьших квадратов)

- **МЕТОДИЧЕСКИЕ УКАЗАНИЯ**

Одной из ключевых предпосылок МНК, обеспечивающих свойство *BLUE*-оценок, является условие постоянства дисперсий случайных отклонений для любых наблюдений. $D(\varepsilon_i)=\sigma^2$ - нарушение предпосылки приводит к изменению свойств оценок, полученных при МНК. Выполнимость данной предпосылки называется *гомоскедастичностью*; невыполнимость данной предпосылки называется *гетероскедастичностью*.

Гетероскедастичность в основном характерна для пространственных или перекрестных данных, реже во временных рядах. Во временных рядах рассматриваются одни и те же показатели в разные моменты времени, поэтому при одновременном росте (или снижении) показателей за определенный период времени может возникнуть гетероскедастичность. При пространственных (перекрестных) данных учитываются различные субъекты, имеющие разные доходы, расходы и т.д.

В качестве примера явной гетероскедастичности можно сказать, что люди с большим доходом не только тратят в среднем больше, чем люди с меньшим доходом, но и разброс в их потреблении также больше, поскольку они имеют больше простора для распределения дохода.

При гетероскедастичности последствия применения МНК будут следующими:

1. Оценки параметров останутся по-прежнему несмещенными и линейными, но перестают быть эффективными (теряют свойство *BLUE*-оценок). Увеличение дисперсии оценок снижает вероятность получения максимально точных оценок.
2. Дисперсии оценок параметров будут рассчитываться со смещением. Поэтому все выводы, получаемые на основе соответствующих *t*- и *F*-статистик, а также интервальные оценки будут ненадежными. Вполне вероятно, что стандартные ошибки коэффициентов будут занижены, а *t*-статистики завышены. Это может привести к признанию статистически значимыми коэффициентов, которые таковыми на самом деле не являются.

Методы выявления гетероскедастичности случайных отклонений. В ряде случаев, зная характер исходных данных, можно предвидеть гетероскедастичность и попытаться устранить проблему ещё на стадии спецификации. Однако чаще всего эту проблему приходится решать после построения уравнения регрессии. Не существует однозначного способа для определения гетероскедастичности.

1. ***Графический способ.*** Графическое построение отклонений от эмпирического уравнения регрессии позволяет визуально определить наличие гетероскедастичности. В этом случае по оси абсцисс откладываются значения объясняющей переменной x_i (для парной регрессии) либо линейную комбинацию объясняющих переменных:

$$\hat{y}_i = a + b_1 x_{i1} + \ldots + b_p x_{ip}, \quad i = \overline{1,n}$$

(для множественной регрессии), а по оси ординат либо отклонения e_i, либо их квадраты $e_t^2, t = \overline{1,n}$.

Если все точки, соответствующие значениям квадратов отклонений e_t^2 находятся внутри горизонтальной полосы постоянной ширины, то это свидетельствует о постоянстве дисперсий e_t^2, т.е. ее независимости от каких либо других факторов – предпосылка о гомоскедастичности случайных отклонений модели регрессии выполняется (рис.1.).

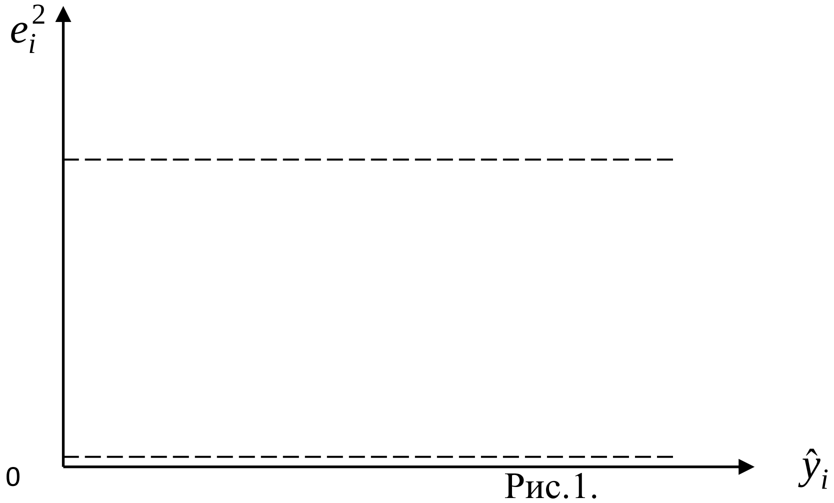


Рис.1.

В других случаях, когда наблюдаются систематические изменения в соотношениях между значениями $\hat{y}_t$ и значениями квадратов отклонений e_t^2 (рис.2 а) и б)), можно говорить о непостоянстве дисперсии отклонений и наличии зависимости между случайными отклонениями и линейной комбинацией экзогенных переменных – предпосылка о гомоскедастичности отклонений модели не выполняется, в модели присутствует гетероскедастичность:

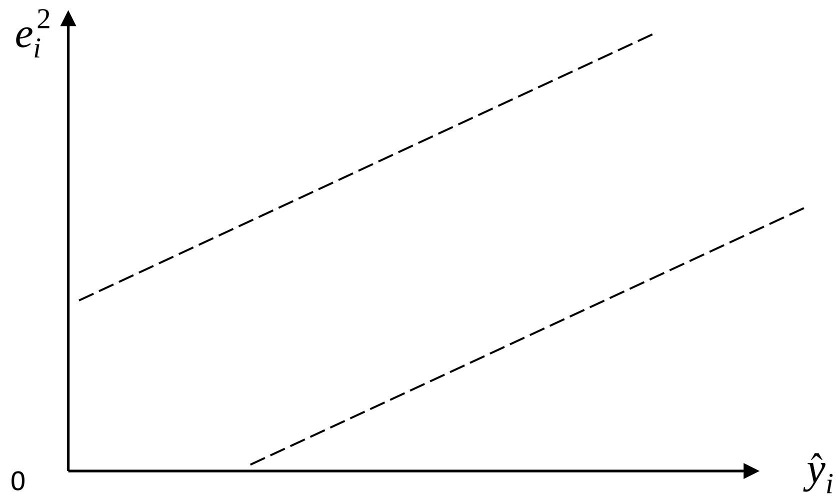


Рис. 2. а)

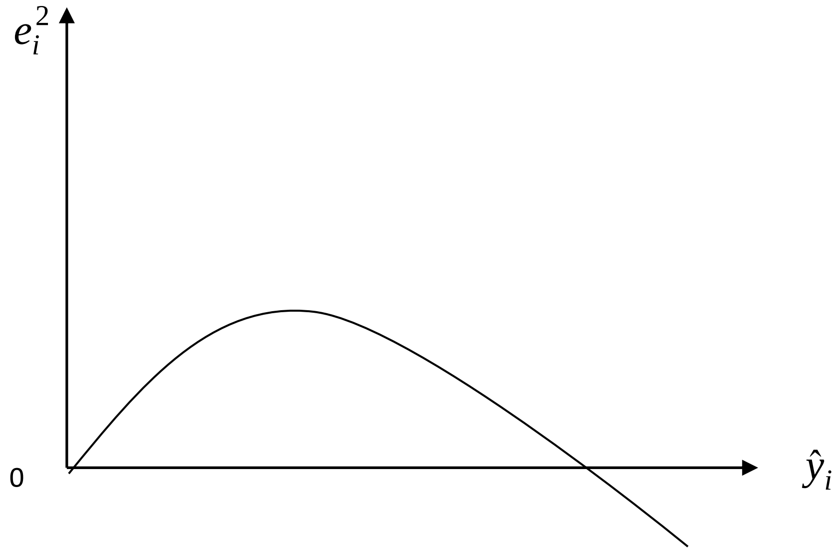


Рис. 2. б)

Графический анализ последних двух графиков отражает ситуации, в которых присутствует большая вероятность наличия гетероскедастичности для рассматриваемых статистических данных. Естественно, что графический анализ должен быть дополнен специальными тестами. В настоящее время для определения гетероскедастичности разработан широкий круг специальных тестов и критериев.

2. ***Тест ранговой корреляции Спирмена.*** В рамках теста предполагается, что дисперсия отклонений будет либо увеличиваться, либо уменьшаться с увеличением значений x_t. Поэтому для регрессии, построенной по МНК, абсолютные величины отклонений $|e_t|$ и значения x_{it} будут в некотором смысле коррелировать (при этом предполагается, что значения экзогенной переменной также положительны). Корреляция в смысле пропорциональности роста абсолютных величин отклонений при росте значений экзогенной переменной приводит нас к понятию ранговой

корреляции: коррелируют между собой не сами значения $e_t|$ и значения x_{it}, а их ранги. Определяется коэффициент ранговой корреляции:

$$r_{e;x} = r\left(rank|e_t|; rank\left(x_{jt}\right)\right) = 1 - 6 \cdot \frac{\sum_{t=1}^{n} d_t^2}{n\left(n^2 - 1\right)},$$

где d_t – разность между рангами x_i и $|e_i|$, n – число наблюдений. Например, если x_{20} является 25-м по величине среди всех значений x_j, а e_{20} является 32-м, то $d_{20} = 25 - 32 = -7$.

Доказано, что если коэффициент корреляции для генеральной совокупности равен нулю , т.е. выполняется гипотеза $H_0 : r_{e;x} = 0$, статистика

$$t_{\text{набл}} = \frac{r_{e;x} \cdot \sqrt{n-2}}{\sqrt{1 - r_{e;x}^2}}$$

имеет распределение Стьюдента с числом степеней свободы (n-2). Поэтому, если наблюдаемое значение t-статистики превышает критическое $t_{\text{крит}} = t_{\alpha/2; n-2}$, вычисленное по таблице критических точек распределения Стьюдента, то гипотезу о равенстве нулю коэффициента корреляции следует отклонить, т.е. признать наличие гетероскедастичности отклонений e_t. В противном случае нулевая гипотеза, которая соответствует отсутствию гетероскедастичности, принимается.

В модели множественной регрессии проверка гипотезы может осуществляться с помощью t – статистики по каждому фактору отдельно. Представленная форма теста может используется в случае несвязанных рангов.

3. ***Тест Голдфелда-Квандта.*** Предполагается, что стандартное отклонение пропорционально значению переменной x_j, т.е. $\sigma_t^2 = \sigma^2 x_{jt}^2$, $t = \overline{1, n}$, а остатки исходной модели e_t имеют нормальное распределение и отсутствует автокорреляция остатков. Далее согласно схеме теста последовательно выполняются:

1. Вся выборка, т.е. входящие в нее переменные, упорядочивается по величине x_t.

2. Упорядоченная выборка разбивается на три подвыборки размерностей k, n-$2k$ и k соответственно. Идея теста состоит в том, что оценки дисперсии отклонений в случае первой и в случае последней подвыборки значительно отличаются в случае невыполнения нулевой гипотезы, т.е. при гетероскедастичности.

3. Для получения оценок дисперсий оцениваются отдельные регрессии для первой подвыборки (k первых наблюдений) и для третьей подвыборки (k последних наблюдений). Поскольку оценка регрессий происходит по выборкам с одинаковым количеством наблюдений, то сравнивать фактически можно значения остаточных сумм квадратов. Если предположение о пропорциональности дисперсий отклонений значениям x_j верно, то

остаточная сумма квадратов отклонений по первой регрессии RSS_1 будет существенно меньше остаточной суммы квадратов отклонений по третьей регрессии RSS_3.

4. Для сравнения соответствующих дисперсий выдвигается нулевая гипотеза в формулировке

$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_n^2$ (случайные отклонения гомоскедастичны)

Для проверки гипотезы строится следующая статистика

$$F_{\text{набл}} = \frac{RSS_3 / (k - m - 1)}{RSS_1 / (k - m - 1)} = \frac{RSS_3}{RSS_1}$$

которая при справедливости нулевой гипотезы имеет распределение Фишера с (*k*-*m*-*1*, *k*-*m*-*1*) степенями свободы. Если $F_{\text{набл}} > F_{\text{крит}} = F_{\alpha; k-m-1; k-m-1}$, то гипотеза об отсутствии гетероскедастичности отклоняется на уровне значимости α.

По рекомендациям специалистов, объем исключаемых данных k должен быть примерно равен четверти общего объёма выборки *n*. Этот же тест может быть использован и при предположении об обратной пропорциональности между дисперсией и значениями объясняющей переменной.

При установлении гетероскедастичности возникает необходимость преобразования модели с целью устранения данного недостатка. Вид преобразования зависит от того, известны или нет дисперсии отклонений σ_i^2.

5. ***Тест Парка.*** Предполагается, что дисперсия σ_t^2 является функцией *t*-го значения экзогенной переменной x_j. Р. Парк предложил следующую функциональную зависимость: $\sigma_t^2 = \sigma^2 \cdot x_t^k \cdot e^{u_t}$, прологарифмировав, получим в линейном виде: $\ln \sigma_t^2 = \ln \sigma^2 + k \cdot \ln x_{jt} + u_t$. Таким образом, задачу проверки предположения о постоянстве дисперсии отклонений σ_t^2 можно свести к проверки значимости зависимости между $\ln \sigma_t^2$ и $\ln x_{jt}$. В случае, если гипотеза о постоянстве дисперсии отклоняется и отклонения гетероскедастичны, параллельно также решается задача определения значения k.

Так как дисперсии σ_t^2 обычно неизвестны, то на практике их заменяют оценками квадратов отклонений e_t^2.

Критерий Парка включает следующие этапы:

- Оценивается исходное уравнение регрессии, например $y_t = \beta_0 + \beta_1 \cdot x_t + \varepsilon_t$, выделяется ряд эмпирических значений остатков e_t ;
- Оценивается уравнение вспомогательной регрессии $\ln e_t^2 = \lambda_0 + \lambda_1 \cdot \ln x_t + u_t$ (в случае рассмотрения множественной регрессии вспомогательная модель теста строится для каждой объясняющей переменной x_j).

- Проверяется статистическая значимость коэффициента λ_1 на основе *t*-статистики. Если коэффициент λ_1 статистически значим, то это означает наличие связи между $\ln e_t^2$ и $\ln x_t$, то есть гетероскедастичность в остатках исходной модели.

К недостаткам теста можно отнести то, что на принятие гипотезы могут влиять свойства отклонений вспомогательной модели u_t, невыполнения предпосылок МНК для которых ведет к искажению результатов (ложно принимается предположение о гетероскедастичности). Кроме того, сформулированная форма зависимости предполагается однозначное определение k, а значит и самой формы, на основе МНК. Т.е. гетероскедастичность может существовать в другой функциональной форме, которая не будет выявлена в тесте Парка.

6. *Тест Глейзера.* Тест Глейзера аналогичен тесту Парка, основывается на более общих представлениях о зависимости стандартной ошибки случайного члена от значений объясняющей переменной, т.е. в каком-то смысле дополняет тест Парка.

Зависимость между случайными отклонениями тестируемой модели и экзогенной переменной представляется в виде $|e_t| = \lambda_0 + \lambda_1 \cdot x_t^l + v_t$. Изменяя значение *l*, можно построить множество регрессий, определяющих разную форму гетероскедастичности. Рекомендуют перебирать значения параметра следующим образом *l=...-2 ,-1, -0,5, 0,5, 1, 2...*. Из общего числа построенных моделей выбирают регрессию с максимальным значением коэффициента детерминации: в общем случае R^2 представляет в тесте функцию от *l*, для которой возможно определение максимального значения при оптимальном значении аргумента *l*. В отобранной модели тестируется статистическая значимость коэффициента λ_1, что фактически означает наличие гетероскедастичности.

Нужно отметить, что также как и в тесте Парка, в тесте Глейзера для отклонений v_t может нарушаться условие гомоскедастичности. Однако во многих случаях предложенные модели являются достаточно хорошими для определения гетероскедастичности. Форма гетероскедастичности в тесте Глейзера может быть определена как $\sigma_t^2 = \sigma^2 \cdot x_t^{2l}$.

7. Тест Уайта. В тесте не накладывается никаких предположений о свойствах случайных отклонений, однако, тест не дает ответа на вопрос о точной форме гетероскедастичности. Для проверки нулевой гипотезы о гомоскедастичности случайных отклонений регрессионной модели необходимо:

- Оценить исходную модель и определить остатки модели e_t;
- Оценить вспомогательную модель регрессии квадратов остатков исходной модели на все ее экзогенные переменные, их квадраты и перекрестные произведения:

$$e_t^2 = \lambda_0 + \lambda_1 \cdot x_{1t} + \lambda_2 \cdot x_{2t} + \ldots + \lambda_m \cdot x_{mt} + \delta_1 \cdot x_{1t}^2 + \delta_2 \cdot x_{2t}^2 + \ldots + \delta_m \cdot x_{mt}^2 +$$
$$+ \alpha_{12} \cdot x_{1t} x_{2t} + \alpha_{13} \cdot x_{1t} x_{3t} + \ldots + \alpha_{1m} \cdot x_{1t} x_{mt} + \alpha_{23} \cdot x_{2t} x_{3t} + \ldots + \alpha_{m-1m} \cdot x_{m-1t} x_{mt} + v_t$$

Определяется R^2 вспомогательного уравнения регрессии и находится значение статистики $Wh = n \cdot R^2 \sim \chi^2_{\alpha;k}$, где k – количество экзогенных факторов во вспомогательной модели (или количество параметров, уменьшенное на единицу). Если $Wh = n \cdot R^2 < \chi^2_{\alpha;k}$, то в остатках модели присутствует гомоскедастичность; в противном случае, если $Wh = n \cdot R^2 > \chi^2_{\alpha;k}$, то нулевая гипотеза отклоняется и в исходной модели присутствует гетероскедастичность.

Возможно использовать форму теста Вайта без перекрестных произведений (no-cross), например, при большом количестве параметров в исходной модели. Так же, поскольку использование статистики χ^2 предполагает большой объем выборки, при недостаточном количестве наблюдений возможно использовать F-статистику, сведя проверку нулевой гипотезы о гомоскедастичности отклонений к проверке гипотезы о статистической незначимости коэффициента детерминации вспомогательной модели.

8. Тест Бреуша-Пагана. Тест, не накладывая никаких ограничений на случайные отклонения, позволяет диагностировать гетероскедастичность в тех случаях, когда она присутствует не только в виде зависимости случайных отклонений от конкретной экзогенной переменной, но и в виде зависимости от нескольких экзогенных переменных, их линейных комбинаций, или когда функция зависимости отличается от степенной (линейной, квадратичной и т.п.), как в предыдущих тестах.

Для проверки нулевой гипотезы о гомоскедастичности случайных отклонений регрессионной модели необходимо:

- Оценить исходную модель и определить остатки модели e_t, а также среднюю величину их квадратов $\hat{\sigma} = \frac{1}{n}\sum_{t=1}^{n} e_t^2$;
- Оценить вспомогательную модель регрессии квадратов остатков исходной модели, деленных на величину $\hat{\sigma}$, на некоторые экзогенные переменные $Z_{1t}, Z_{2t}, \ldots$, получаемые использованием элементарных алгебраических функций или различных комбинаций из экзогенных переменных исходной модели:

$$\frac{e_t^2}{\hat{\sigma}} = \lambda_0 + \lambda_1 \cdot Z_{1t} + \lambda_2 \cdot Z_{2t} + \ldots + \lambda_m \cdot Z_{mt} + v_t$$

Определяется значение объясненной суммы квадратов ESS вспомогательного уравнения регрессии и находится значение статистики $BP = ESS/2 \sim \chi^2_{\alpha;k}$, где k – количество экзогенных факторов во вспомогательной модели (или количество параметров, уменьшенное на единицу). Если $BP < \chi^2_{\alpha;k}$, то в остатках модели присутствует гомоскедастичность; в противном случае, если $BP > \chi^2_{\alpha;k}$, то нулевая гипотеза отклоняется и в исходной модели присутствует гетероскедастичность. Можно использовать также значение остаточной суммы RSS вспомогательной модели, выводы при этом меняются на противоположные.

Задание 2.2.1

Исследовался вопрос зависимости расходов на научную и исследовательскую деятельность (переменная RnD – research and development expenditure) от объема прибыли (переменная Profits) в 18 промышленных группах, данные представлены в таблице 2.7.

Таблица 2.7

Номер группы	Profits	R&D Expenditure	Номер группы	Profits	R&D Expenditure
1	185,1	62,5	10	13869,9	6620,1
2	1569,5	92,9	11	4487,8	3918,6
3	276,8	178,3	12	10278,9	1595,3
4	2828,1	258,4	13	8787,3	6107,5
5	225,9	494,7	14	16438,8	4454,1
6	3751,9	1083	15	9761,4	3163,8
7	2884,1	1620,6	16	19774,5	13210,7
8	4645,7	421,7	17	22626,6	1703,8
9	5036,4	509,2	18	18415,4	9528,2

Источник: Gujarati N. Damodar: Basic Econometrics, 5th ed, McGraw Hill Higher Education, 2009.

Оцените модель линейной регрессии переменной R&D на переменную объема прибыли. Проанализируйте отсутствие зависимости между случайными отклонениями модели и экзогенной переменной, сформулировав предположение в виде гипотезы о гомоскедастичности случайных отклонений регрессионной модели и проверив ее выполнение несколькими способами. В случае, если в результате анализа гипотеза будет отклонена, скорректируйте гетероскедастичность с помощью взвешенного метода наименьших квадратов.

Используя принятые в таблице обозначения, оценим параметры регрессии $\widehat{RnD}_t = \beta_0 + \beta_1 \cdot \mathrm{Pr}ofits_t$:

Модель 1

Зависимая (эндогенная) переменная *RnD*

Количество наблюдений 18

Переменная	Коэффициент	Ст. ошибка	t-стат.	P-вероятн.
Константа *C*	114,3906	959,0376	0,1193	0,9065
Независимая (экзогенная) переменная *Profits*	0,3632	0,0892	4,0735	0,0009
Коэффициент детерминации	0,509102	F-статистика		16,5933

Полученные результаты указывают на статистическую значимость коэффициента при экзогенной переменной и статистическую значимость коэффициента детерминации R^2, не смотря на относительно небольшое абсолютное значение (значение соответствующей статистики $F_{набл} = 16{,}5933$ превышает значение критической точки распределения Фишера $F_{крит} = F(0{,}05;1;16) = 4{,}494$). По формальным признакам нет оснований предположить невыполнение для случайных отклонений модели (1) какой-либо из предпосылок МНК, поэтому принимаем решение провести дополнительно анализ отклонений модели с помощью теста Голдфелда-Квандта, а также с помощью теста Вайта, считающегося универсальным, поскольку процедура не имеет ограничений на структуру гетероскедастичности.

Тест Голдфельда-Квандта (Goldfeld-Quandt). Прежде всего проранжируем выборку по значениям экзогенной переменной *Profits* и исключим из выборки примерно четверть центральных наблюдений (получив *18:4=4,5*, для удобства округляем полученное число в меньшую сторону, чтобы получить одинаковое количество наблюдений со значениями меньше и больше исключенных: $n' = 7$ – объем подвыборок):

Номер группы	Profits	R&D Expenditure	Номер группы	Profits	R&D Expenditure
1	185,1	62,5	***9***	***5036,4***	***509,2***
5	225,9	494,7	***13***	***8787,3***	***6107,5***
3	276,8	178,3	15	9761,4	3163,8
2	1569,5	92,9	12	10278,9	1595,3
4	2828,1	258,4	10	13869,9	6620,1
7	2884,1	1620,6	14	16438,8	4454,1
6	3751,9	1083	18	18415,4	9528,2
11	***4487,8***	***3918,6***	16	19774,5	13210,7
8	***4645,7***	***421,7***	17	22626,6	1703,8

Для двух полученных подвыборок, в каждой из которых содержится 7 наблюдений, оценивается модель регрессии со спецификацией аналогичной исходной модели и выписывается значение остаточной суммы:

Тест Голдфельда-Квандта (Goldfeld-Quandt):

Вспомогательное регрессионное уравнение (1я подвыборка)

Зависимая (эндогенная) переменная *RnD*

Количество наблюдений 7 (с 1 по 7 наблюдение в проранжированной выборке)

Переменная	Коэффициент	Ст. ошибка	t-стат.	Р-вероятн.

C	111,3779	293,0639	0,3800	0,7195
Profits	0,2569	0,1349	1,9035	0,1153
Коэффициент детерминации	0,420186	Сумма квадратов отклонений		1219124,77

Вспомогательное регрессионное уравнение (2я подвыборка)

Зависимая (эндогенная) переменная *RnD*

Количество наблюдений 7 (с 12 по 18 наблюдение в проранжированной выборке)

Переменная	Коэффициент	Ст. ошибка	t-стат.	P-вероятн.
C	324,2195	6113,783	0,0530	0,9598
Profits	0,3419	0,3705	0,9227	0,3985
Коэффициент детерминации	0,145495	Сумма квадратов отклонений		96298297,6339

Для проверки гипотезы о гомоскедастичности случайных отклонений модели (1) с помощью теста Голдфелда-Квандта находим F-статистику, на основе которой проверяем гипотезу о равенстве остаточных сумм двух регрессионных моделей, оцененных по данным подвыборок (т.е. фактически гипотезу о гомоскедастичности отклонений модели, построенной наблюдениям полной выборки) :

$$H_0 : RSS_1 = RSS_2 \Leftrightarrow \text{гомоскедастичность отклонений модели (1)}$$

$$H_0 : RSS_1 < RSS_2 \Leftrightarrow \text{гетероскедастичность отклонений модели (1)}$$

$$F_{набл} = \frac{RSS_2 /(n' - m - 1)}{RSS_1 /(n' - m - 1)} = \frac{RSS_2}{RSS_1} \sim F(\alpha\, ; n' - m - 1; n' - m - 1)$$

$$F_{набл} = \frac{RSS_2}{RSS_1} = \frac{96298297{,}6339}{1219124{,}77} = 78{,}9897 \sim F_{крит} = F(0{,}05;5;5) = 5{,}05$$

Т.к. $F_{набл} = 78{,}9897 > F_{крит} = 5{,}05$, то нулевая гипотеза отклоняется в пользу альтернативной и подтверждается предположение о гетероскедастичности отклонений модели (1).

Тест Вайта (White). Для проверки гипотезы о гомоскедастичности случайных отклонений исходной модели построим вспомогательную модель вида $e_t^{\,2} = \gamma_0 + \gamma_1 \cdot Profits_t + \gamma_2 \cdot Profits_t^{\,2} + u_t$:

Тест Вайта (White) :

H_0 : гомоскедастичность случайных отклонений модели (по переменной *Profits*)

H_1 : гетероскедастичность случайных отклонений модели

F-статистика для R^2	23,8985	$F_{0,05}(2{,}15)$ =3,682	Р-вероятность	0,0000
Статистика $n \cdot R^2$	13,7004	$\chi^2_{0,05}(2) = 5{,}99$	Р-вероятность	0,0011

Вспомогательное регрессионное уравнение

Зависимая (эндогенная) переменная $e_t^{\,2}$

Количество наблюдений 18

Переменная	Коэффициент	Ст. ошибка	t-стат.	Р-вероятн.
C	3514555	3062699	1,1475	0,2691
Profits	-1582,81	806,5255	-1,9625	0,0685
Profits ^{2}	0,1355	0,0371	3,6564	0,0023
Коэффициент детерминации	0,761135	F-статистика		23,8985

Для проверки гипотезы о гомоскедастичности случайных отклонений модели (1) с помощью теста Вайта используем значение статистики $Wh = n \cdot R^2$, где *n* - число наблюдений, *R*^{2}– коэффициент детерминации вспомогательной модели. С помощью таблицы критических значений χ^2-распределения находим критическую точку $\chi^2_{0,05}(2) = 5{,}99$. Поскольку значение статистики $Wh = n \cdot R^2 = 13{,}7004 > 5{,}99 = \chi^2_{0,05}(2)$, то *H₀* отклоняется в пользу альтернативной гипотезы о гетероскедастичности случайных отклонений модели (1).

Тест Вайта предполагает также возможность использования F-статистики гипотезы о статистической значимости коэффициента детерминации (рекомендуется, например, при малых выборках, поскольку в общем случае тест Вайта асимптотический). Выводы, получаемые на основе F-статистики, как правило, согласуются с выводами, принимаемыми при использовании статистики $Wh = n \cdot R^2$ и критических значений χ^2-распределения, поскольку соответствующие значения P-вероятностей принятия нулевой гипотезы близки. Для проверки гипотезы находим значение критической точки в таблице распределения Фишера для уровня значимости α=0,05 и значений степеней свободы $\nu_1 = m = 2, \nu_2 = n - m - 1 = 15$: $F_{крит} = 3{,}682$. Т.к. $F_{набл} = 23{,}8985 > F_{крит}$, то нулевая гипотеза отклоняется в пользу предположения о гетероскедастичности отклонений.

Тест Парка (Park). Для подтверждения выводов, полученных при проверке гипотезы о гомоскедастичности случайных отклонений с помощью теста Вайта, и для выявления формы гетероскедастичности, проведем тест Парка. В рамках процедуры теста для случайных отклонений исходной модели построим вспомогательную модель вида $\ln e_t^2 = \gamma_0 + \gamma_1 \cdot \ln Profits_t + u_t$:

ТестParka (Park) :

H_0 : гомоскедастичность случайных отклонений модели (по переменной $Profits_t$)

H_1 : гетероскедастичность случайных отклонений модели

Вспомогательное регрессионное уравнение

Зависимая (эндогенная) переменная $\ln e_t^{\,2}$

Количество наблюдений 18

Переменная	Коэффициент	Ст. ошибка	t-стат.	P-вероятн.
C	1,3003	1,8515	0,7023	0,4926
ln Profits	1,4940	0,2192	6,8142	0,0000
Коэффициент детерминации	0,743725	F-статистика		46,433

Для проверки гипотезы о гомоскедастичности случайных отклонений модели (1) на основе теста Парка используем значение статистики t-статистики коэффициента при переменной *lnProfits* вспомогательной модели. Используем критическую точку распределения Стьюдента при уровне значимости α=0,05 и значении степеней свободы ν=n-m-1=18-1-1=16: $t_{\text{крит}} = t_{\alpha/2; n-m-1} = 2{,}12$. Поскольку значение статистики $|t| = 6{,}8142 > 2{,}12 = t_{\text{крит}} = t_{\alpha/2; n-m-1}$, то H_0 должна быть отклонена в пользу H_1 при выбранном уровне значимости. Следовательно, предпосылка Гаусса-Маркова о гомоскедастичности случайных отклонений модели нарушается. Р-вероятность для t-статистики показывает, что гипотеза H_0 отклоняется и при уровне значимости α=0,01, т.к. α>P=0,0000. Форма гетероскедастичности согласно результатам теста имеет вид: $e_t^2 = \sigma^2 \cdot Profits_t^{\,1{,}5}$ (с учетом округления).

В случае, когда в тесте Парка нулевая гипотеза отклоняется, т.е. коэффициент γ_1 вспомогательной модели статистически значим, необходимо проверить выполнение предпосылок МНК для случайных отклонений u_t вспомогательной модели. Случайные отклонения u_t могут оказаться коррелированны или гетероскедастичны, что приводит к смещению дисперсии коэффициента γ_1, получению завышенного значения соответствующей t-статистики и ошибке при проверке гипотезы. В нашем случае случайные отклонений u_t удовлетворяют предпосылкам МНК, что можно проверить, например, с помощью графического анализа, теста Вайта и т.д.

Оценка параметров модели с помощью ВМНК. Согласно результатам тестирования в исходной модели (1) диагностирована гетероскедастичность случайных отклонений, что приводит нас к необходимости исправления построенной модели методом взвешенных наименьших квадратов, в котором при определении веса используем результаты, полученные в тесте Парка при определении формы гетероскедастичности. Поскольку дисперсия отклонений пропорциональна значениям экзогенной переменной *Profits*$_t$ в степени *k=1,5*, то значения случайных отклонений пропорциональны значениям экзогенной переменной $\Pr ofits_t$ в степени *k:2=1,5:2=0,75*, следовательно вся модель будет «взвешена» на эти значения:

МНК: $\hat{RnD_t} = \beta_0 + \beta_1 \cdot \Pr ofits_t$

ВМНК: $\left\langle \frac{RnD_t}{\Pr ofits_t^{3/4}} \right\rangle = \rho_0 \cdot \left\langle \frac{1}{\Pr ofits_t^{3/4}} \right\rangle + \rho_1 \cdot \left\langle \frac{\Pr ofits_t}{\Pr ofits_t^{3/4}} \right\rangle + \xi_t$, где $\xi_t = \frac{e_t}{\Pr ofits_t^{3/4}} = \sqrt{\sigma^2} = const$ полагаются гомоскедастичными. В общем случае, как можно заметить, модель ВМНК представляет собой модель множественной линейной регрессии без константы (исключение составляет случай, когда *k=2*). Полученные при оценке модели ВМНК значения параметров и числовых характеристик в случае подтверждения гомоскедастичности отклонений ξ_t переносятся в исходную спецификацию модели:

Модель 2

Зависимая (эндогенная) переменная *RnD**

Количество наблюдений 18

Переменная	Коэффициент	Ст. ошибка	t-стат.	P-вероятн.
Константа *C**	139,8665	92,9887	1,5041	0,1520
Независимая (экзогенная) переменная	0,3506	0,0762	4,6016	0,0003

*Profits**

Р-вероятность в тесте Вайта (F-ст.)	0,7272	Переоцененный R^2	0,508036

Задание 2.2.2

Исследовался вопрос зависимости расходов на научную и исследовательскую деятельность (переменная RnD – research and development expenditure) от уровня продаж (переменная Sales) в 18 промышленных группах, данные представлены в таблице:

Таблица 2.8

Номер группы	Sales	R&D Expenditure	Номер группы	Sales	R&D Expenditure
1	6375,3	62,5	10	80552,8	6620,1
2	11626,4	92,9	11	95294,0	3918,6
3	14655,1	178,3	12	101314,1	1595,3
4	21869,2	258,4	13	116141,3	6107,5
5	26408,3	494,7	14	122315,7	4454,1
6	32405,6	1083	15	141649,9	3163,8
7	35107,7	1620,6	16	175025,8	13210,7
8	40295,4	421,7	17	230614,5	1703,8
9	70761,6	509,2	18	293543	9528,2

Источник: Gujarati N. Damodar: Basic Econometrics, 5th ed, McGraw Hill Higher Education, 2009.

Оцените модель линейной регрессии переменной R&D на переменную уровня продаж. Проанализируйте отсутствие зависимости между случайными отклонениями модели и экзогенной переменной, сформулировав предположение в виде гипотезы о гомоскедастичности случайных отклонений регрессионной модели и проверив ее выполнение несколькими способами. В случае, если в результате анализа гипотеза будет отклонена, скорректируйте гетероскедастичность с помощью взвешенного метода наименьших квадратов.

Используя принятые в таблице обозначения, оценим параметры регрессии $\widehat{RnD}_t = \beta_0 + \beta_1 \cdot Sales_t$:

Модель 1

Зависимая (эндогенная) переменная *RnD*

Количество наблюдений 18

Переменная	Коэффициент	Ст. ошибка	t-стат.	P-вероятн.
Константа *C*	192,9931	990,9858	0,1947	0,8480
Независимая (экзогенная) переменная *Sales*	0,0319	0,0083	3,8300	0,0015
Коэффициент детерминации	0,478303	F-статистика		14,6692

Полученные результаты указывают на статистическую значимость коэффициента при экзогенной переменной и статистическую значимость коэффициента детерминации R^2, не смотря на относительно небольшое абсолютное значение (значение соответствующей статистики $F_{набл} = 14,6692$ превышает значение критической точки распределения Фишера $F_{крит} = F(0,05;1;16) = 4,494$). По формальным признакам нет оснований предположить невыполнение для случайных отклонений модели (1) какой-либо из предпосылок МНК, поэтому принимаем решение провести анализ отклонений модели с помощью теста Вайта, считающимся универсальным, поскольку процедура не имеет ограничений на структуру гетероскедастичности.

Тест Вайта (White). Для проверки гипотезы о гомоскедастичности случайных отклонений исходной модели построим вспомогательную модель вида $e_t^2 = \gamma_0 + \gamma_1 \cdot Sales_t + \gamma_2 \cdot Sales_t^2 + u_t$:

Тест Вайта (White) :

H_0 : гомоскедастичность случайных отклонений модели

H_1 : гетероскедастичность случайных отклонений модели

F-статистика для R^2	3,0572	$F_{0,05}(2,15)$ =3,682	P-вероятность	0,0770
Статистика $n \cdot R^2$	5,2125	$\chi^2{}_{0,05}(2) = 5{,}99$	P-вероятность	0,0738

Вспомогательное регрессионное уравнение

Зависимая (эндогенная) переменная $e_t^{\,2}$

Количество наблюдений 18

Переменная	Коэффициент	Ст. ошибка	t-стат.	P-вероятн.
C	-6219665	6459809	-0,9628	0,3509
Sales	229,3508	126,2197	1,8171	0,0892
Sales ^{2}	-0,0005	0,0004	-1,1950	0,2507
Коэффициент детерминации	0,289583	F-статистика		3,057178

Для проверки гипотезы о гомоскедастичности случайных отклонений модели (1) с помощью теста Вайта используем значение статистики $Wh = n \cdot R^2$, где *n* – число наблюдений, *R*^{2}– коэффициент детерминации вспомогательной модели. С помощью таблицы критических значений χ^2-распределения находим критическую точку $\chi^2{}_{0,05}(2) = 5{,}99$. Поскольку значение статистики $Wh = n \cdot R^2 = 5{,}2125 < 5{,}99 = \chi^2{}_{0,05}(2)$, то H_0 не отклоняется при этом уровне значимости. Следовательно, предпосылка Гаусса-Маркова о гомоскедастичности случайных отклонений модели выполняется. С другой стороны, P-вероятность для статистики Wh показывает, что гипотеза H_0 отклоняется при уровне значимости α=0,10, т.к. в этом случае α>P=0,0738. Можно сделать вывод о том, что в модели присутствует *слабая* гетероскедастичность отклонений.

При использовании F-статистики выводы о принятии или отклонении нулевой гипотезы о гомоскедастичности отклонений модели (1) также различаются в зависимости от уровня значимости: находим значение критической точки в таблице распределения Фишера для уровня значимости α=0,05 и значений степеней свободы $\nu_1 = m = 2,\ \nu_2 = n - m - 1 = 15$: $F_{крит} = 3{,}682$; а также при α=0,10: $F_{крит} = 2{,}695$. При α=0,05 $F_{набл} < F_{крит}$ и нулевая (основная) гипотеза принимается, но при α=0,10 $F_{набл} > F_{крит}$ – нулевая гипотеза отклоняется в пользу предположения о гетероскедастичности отклонений. Исходя из точного значения P-вероятности, коэффициент детерминации вспомогательной модели регрессии в тесте Вайта статистически значим для α≥0,08, следовательно, при соответствующем уровне значимости существует статистически значимая зависимость между дисперсией отклонений модели (1) и значениями экзогенной переменной *Sales* и случайные отклонения исходной модели не обладают желаемым свойством постоянства дисперсии, т.е. гомоскедастичностью.

Тест Парка (Park). Для подтверждения выводов, полученных при проверке гипотезы о гомоскедастичности случайных отклонений с помощью теста Вайта, и для выявления формы гетероскедастичности, проведем тест Парка. В рамках процедуры теста для случайных отклонений исходной модели построим вспомогательную модель вида $\ln e_t^2 = \gamma_0 + \gamma_1 \cdot \ln Sales_t + u_t$:

Тест Парка (Park) :

H_0 : гомоскедастичность случайных отклонений модели (по переменной $Sales_t$)

H_1 : гетероскедастичность случайных отклонений модели

Вспомогательное регрессионное уравнение

Зависимая (эндогенная) переменная $\ln e_t^2$

Количество наблюдений 18

Переменная	Коэффициент	Ст. ошибка	t-стат.	P-вероятн.
C	5,6877	6,6352	0,8572	0,4040
lnSales	0,7014	0,6033	1,1626	0,2620
Коэффициент детерминации	0,077897	F-статистика		1,3516

Значение Р-вероятности для t-статистики коэффициентов вспомогательной модели показывает, что гипотеза H_0 о гомоскедастичности отклонений исходной модели принимается, т.к. коэффициенты регрессии статистически незначимы, как и коэффициент детерминации, а значит не существует взаимосвязи между отклонениями e_t и экзогенной переменной *Sales*.

Полученный в тесте Парка вывод противоречит результатам теста Вайта, где был найден допустимый уровень значимости α, при котором можно отклонить нулевую гипотезу и принять верным предположение о гетероскедастичности случайных отклонений. В данном случае проявляет себя некоторое ограничение теста Парка, поскольку в тесте предполагается наличие зависимости вида $e_t^2 = \sigma^2 \cdot Sales_t^k$, где значение *k* определяется с помощью МНК, т.е. однозначно. В этом смысле дополняющим тест Парка можно считать тест Глейзера, поскольку он позволяет подобрать значение *k* перебором из нескольких значений, оцениваемых также на основе МНК.

Тест Глейзера (Glaser). Для проверки гипотезы о гомоскедастичности случайных отклонений исходной модели предполагается построение вспомогательной модели вида $|e_t| = \gamma_0 + \gamma_1 \cdot Sales_t^p + u_t$ при различных значениях параметра *p*. В общем случае значения *p* перебирают, т.е. $p = \{\pm 0{,}25; \pm 0{,}5; \pm 1; \pm 2 \ldots\}$, однако можно сузить интервал исследования, если исходить из того, что оптимальное значение *p* находится близко на числовой прямой к значению $k/2$, получаемому в тесте Парка и равному в данном задании 0,3. Как правило, при переборе *p* либо определяется единственное его значение, при котором вспомогательная модель имеет максимальное значение коэффициента детерминации R^2, либо значение коэффициента детерминации с увеличением (уменьшением) *p* постоянно растет, приближаясь к единице, при этом достигаются настолько большие или малые значения *p*, что процедура теряет смысл.

Тест Глейзера (Glaser):

Вспомогательное регрессионное уравнение $p = \frac{1}{5}$

Зависимая (эндогенная) переменная $|e_t|$

Количество наблюдений 18

Переменная	Коэффициент	Ст. ошибка	t-стат.	P-вероятн.
C	-3477,191	2175,679	-1,5982	0,1296
Sales	561,9899	233,7332	2,4044	0,0287
Коэффициент детерминации	0,265421			

Вспомогательное регрессионное уравнение $p = \frac{1}{4}$

Зависимая (эндогенная) переменная $|e_t|$

Количество наблюдений 18

Переменная	Коэффициент	Ст. ошибка	t-стат.	P-вероятн.
C	-2498,200	1776,2050	-1,4065	0,1787
Sales	259,9411	107,9705	2,4075	0,0285
Коэффициент детерминации	0,265926			

Вспомогательное регрессионное уравнение $p = \frac{1}{3}$

Зависимая (эндогенная) переменная $|e_t|$

Количество наблюдений 18

Переменная	Коэффициент	Ст. ошибка	t-стат.	P-вероятн.
C	-1512,158	1383,935	-1,0927	0,2907
Sales	77,6156	32,2765	2,4047	0,0286
Коэффициент детерминации	0,265469			

Выбираем значение степени *p* при котором вспомогательная модель имеет максимальное значение коэффициента детерминации: $p = 1/4$ и $R^2=0,265926$. Значение *P*-вероятности для *t*-статистики коэффициента γ_1 выбранной вспомогательной модели указывает на его статистическую значимость, что подтверждает вывод о том, что гипотеза H_0 о гомоскедастичности отклонений исходной модели может быть отклонена.

Оценка параметров модели с помощью ВМНК. Согласно результатам тестирования в исходной модели (1) диагностирована *слабая* гетероскедастичность случайных отклонений. При выборе 5% уровня значимости для проведения исследования нет необходимости корректировать модель на гетероскедастичность. Применим к модели (1) метод взвешенных наименьших квадратов с целью показать корректность определения формы гетероскедастичности в тесте Глейзера. Поскольку модули случайных отклонений пропорциональны значениям экзогенной переменной $Sales_t$ степени *p=0,25*, то в методе ВМНК вся модель будет «взвешиваться» на эти значения:

МНК: $\widehat{RnD}_t = \beta_0 + \beta_1 \cdot Sales_t$

ВМНК: $\left\langle \dfrac{RnD_t}{Sales_t^{1/4}} \right\rangle = \rho_0 \cdot \left\langle \dfrac{1}{Sales_t^{1/4}} \right\rangle + \rho_1 \cdot \left\langle \dfrac{Sales_t}{Sales_t^{1/4}} \right\rangle + \xi_t$,

где $\xi_t = \dfrac{e_t}{Sales_t^{1/4}} = \sqrt{\sigma^2} = const$ полагаются гомоскедастичными.

В общем случае, как можно заметить, модель ВМНК представляет собой модель множественной линейной регрессии без константы (исключение составляет случай, когда $k = 2$). Полученные при оценке модели ВМНК значения параметров и числовых характеристик в случае подтверждения гомоскедастичности отклонений ξ_t переносятся в исходную спецификацию модели:

Модель 2

Зависимая (эндогенная) переменная *RnD**				
Количество наблюдений 18				
Переменная	Коэффициент	Ст. ошибка	t-стат.	P-вероятн.
Константа *C**	-119,2569	635,5568	-0,1876	0,8535
Независимая (экзогенная) переменная *Sales**	0,0349	0,0076	4,5986	0,0003
P-вероятность в тесте Вайта (F-ст.)	0,1370	Переоцененный R^2		0,474024

Задание 2.2.3

В задаче используются статистические данные из примера 2.1.1 (таблица 2.1). Проверьте для полученной в примере линейной регрессионной зависимости потребительских расходов от объема валового продукта $\hat{CN}_t = \beta_0 + \beta_1 GDP_t$ гипотезу о гомоскедастичности случайных отклонений:

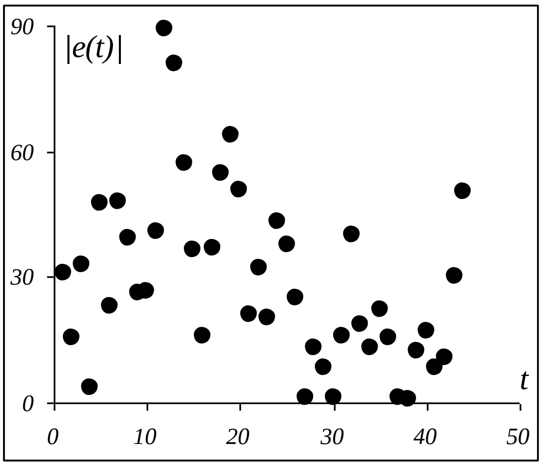
Модель 1

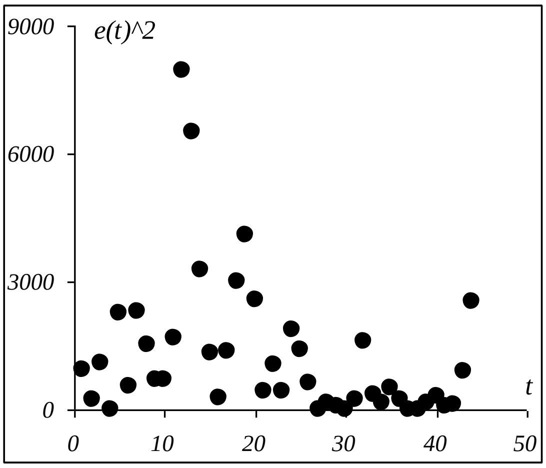
Зависимая (эндогенная) переменная *CN*
Количество наблюдений 44

Переменная	Коэффициент	Ст. ошибка	t-стат.	Р-вероятн.
Константа С	-54,24070	46,86539	-1,157372	0,2537
Независимая (экзогенная) переменная *GDP*	0,663393	0,008719	76,08230	0,0000
Коэффициент детерминации	0,992797	F-статистика		5788,516

Полученные результаты указывают на статистическую значимость коэффициента при экзогенной переменной; в тоже время, исходя из чрезмерно высоких значений коэффициента детерминации и t-статистики, по формальным признакам можно сделать предварительный вывод о наличии как гетероскедастичности, так и коррелированности случайных отклонений модели (что и было подтверждено в примере 2.1.1). В этом случае, полученные значения дисперсий смещены (стандартная ошибка регрессии, стандартные ошибки коэффициентов) и нами могут быть сделаны ложные выводы о качестве построенной модели.

Графический метод. Визуальный анализ разброса на декартовой плоскости точек, соответствующих значениям случайных отклонений e_t (или абсолютных значений $|e_t|$, или квадратов значений e_t^2) позволяет предварительно сделать вывод о постоянстве дисперсии отклонений модели. В общем случае, рекомендуется построить все предлагаемых варианты, поскольку наглядность графика разброса будет зависеть непосредственно от фактических числовых значений отклонений:





Так, в нашем случае при анализе графика разброса абсолютных значений случайных отклонений достаточно сложно сделать вывод о гетероскедастичности, в отличие от графика разброса квадратов отклонений, на котором визуально легко заметить уменьшение разброса значений с увеличением номера наблюдения *t*. Таким образом, графический анализ позволяет сделать вывод о наличии гетероскедастичности случайных отклонений исходной модели и дает основания для дальнейшего тестирования.

Тест Парка (Park). Для проверки гипотезы о гомоскедастичности случайных отклонений исходной модели построим вспомогательную модель вида $\ln e_t^2 = \gamma_0 + \gamma_1 \cdot \ln GDP_t + u_t$:

Тест Рарка (Park) :

H_0 : гомоскедастичность случайных отклонений модели (по переменной GDP_t)

H_1 : гетероскедастичность случайных отклонений модели

Вспомогательное регрессионное уравнение

Зависимая (эндогенная) переменная $\ln e_t^2$

Количество наблюдений 44

Переменная	Коэффициент	Ст. ошибка	t-стат.	P-вероятн.

C	74,9051	22,7817	3,2880	0,0020
lnGDP	-8,0426	2,6563	-3,0277	0,0042
Коэффициент детерминации	0,179157	F-статистика		9,1669

Для проверки гипотезы о гомоскедастичности случайных отклонений модели (1) на основе теста Парка используем значение статистики t-статистики коэффициента при переменной *lnGDP* вспомогательной модели. Используем критическую точку распределения Стьюдента при уровне значимости α=0,05 и значении степеней свободы $\nu = n - m - 1 = 44 - 1 - 1 = 42$: $t_{крит} = t_{\alpha/2; n-m-1} = 2{,}018$. Поскольку значение статистики $|t| = 3{,}028 > 2{,}018 = t_{крит} = t_{\alpha/2; n-m-1}$, то H_0 должна быть отклонена в пользу H_1 при выбранном уровне значимости. Следовательно, предпосылка Гаусса-Маркова о гомоскедастичности случайных отклонений модели нарушается. Р-вероятность для t-статистики показывает, что гипотеза H_0 отклоняется и при уровне значимости α=0,01, т.к. *α>P=0,0042*. Форма гетероскедастичности согласно результатам теста имеет вид: $e_t^2 = \sigma^2 \cdot GDP_t^{-8}$.

В случае, когда в тесте Парка нулевая гипотеза отклоняется, т.е. коэффициент γ_1 вспомогательной модели статистически значим, необходимо проверить выполнение предпосылок МНК для случайных отклонений u_t вспомогательной модели. Случайные отклонения u_t могут оказаться коррелированны или гетероскедастичны, что приводит к смещению дисперсии коэффициента γ_1, получению завышенного значения соответствующей t-статистики и ошибке при проверке гипотезы.

Тест Вайта (White). Для проверки гипотезы о гомоскедастичности случайных отклонений исходной модели построим вспомогательную модель вида $e_t^2 = \gamma_0 + \gamma_1 \cdot GDP_t + \gamma_2 \cdot GDP_t^2 + u_t$:

Тест Вайта (White) :

H_0 : гомоскедастичность случайных отклонений модели

H_1 : гетероскедастичность случайных отклонений модели

F-статистика для R^2	4,0996	$F_{0,05}(2,41)$ =3,226	P-вероятность	0,0238
Статистика $n \cdot R^2$	7,3327	$\chi^2{}_{0,05}(2) = 5,99$	P-вероятность	0,0256

Вспомогательное регрессионное уравнение

Зависимая (эндогенная) переменная e_t^2

Количество наблюдений 44

Переменная	Коэффициент	Ст. ошибка	t-стат.	P-вероятн.
C	-984,7029	23280,2	-0,0423	0,9665
GDP	1,95099	8,84610	0,2206	0,8265
GDP ^{2}	-0,00028	0,00083	-0,3406	0,7352
Коэффициент детерминации	0,166652	F-статистика		4,099567

Находим значение статистики $Wh = n \cdot R^2$, где *n* – число наблюдений, R^2 - коэффициент детерминации вспомогательной модели. Поскольку значение статистики $Wh = n \cdot R^2 = 7{,}3327 > 5{,}99 = \chi^2{}_{0,05}(2)$, то H_0 должна быть отклонена в пользу H_1 при этом уровне значимости. Следовательно, предпосылка Гаусса-Маркова о гомоскедастичности случайных отклонений модели нарушается. P-вероятность для статистики *Wh* показывает, что гипотеза H_0 не отклоняется при уровне значимости α=0,02, т.к. α<P=0,0256. Можно сделать вывод о том, что в модели присутствует гетероскедастичность отклонений.

При использовании F-статистики нулевая гипотеза также отклоняется в пользу альтернативной о гетероскедастичности отклонений модели (1): находим значение критической точки в таблице распределения Фишера для уровня значимости α=0,05 и значений степеней свободы $\nu_1 = m = 2,\ \nu_2 = n - m - 1 = 41$: $F_{крит} = 3{,}226$. Поскольку $F_{набл} > F_{крит}$, то нулевая (основная) гипотеза отклоняется, т.е. коэффициент детерминации вспомогательной модели регрессии в тесте Вайта статистически значим, следовательно, существует статистически значимая зависимость между дисперсией отклонений модели (1) и значениями экзогенной переменной GDP и случайные отклонения исходной модели не обладают желаемым свойством постоянства дисперсии, т.е. гомоскедастичностью.

Оценка регрессионных моделей в условиях нарушения предпосылок МНК. Мультиколлинеарность: понятие и последствия, методы выявления, вопросы, связанные с исправлением.

Ключевые понятия: мультиколлинеарность, парные и частные коэффициенты корреляции, метод инфляционных факторов.

- **МЕТОДИЧЕСКИЕ УКАЗАНИЯ**

Одним из условий Гаусса-Маркова является условие об отсутствии строгой (сильной) линейной зависимости между экзогенными переменными. Это условие предполагает, что столбцы матрицы регрессоров X являются независимыми и матрица $\left(X^T X\right)^{-1}$ имеет ранг $r = m+1$, где m – количество экзогенных факторов. При нарушении данного предположения говорят, что в регрессионной модели присутствует мультиколлинеарность.

Проблема мультиколлинеарности является обычной для регрессий временных рядов, т.к. данные состоят из ряда наблюдений в течение какого-то периода времени. Так, объясняющие переменные могут иметь общий временной тренд, относительно которого они совершают малые колебания. Значение одних независимых переменных являются лагированными значениями других. Под совершенной (явной, полной) мультиколлинеарностью понимается существование между некоторыми из факторов линейной функциональной связи. Несовершенная (частичная, неявная) мультиколлинеарность возникает в случаях существования достаточно тесных линейных статистических связей между объясняющими переменными. Важно выявлять те ситуации, когда мультиколлинеарность является следствием неправильной спецификации регрессионной модели.

Последствия мультиколлинеарности:

1. Оценки параметров регрессии смещены, возможно получение неверного знака у коэффициентов регрессии;
2. Дисперсии оценок вычисляются со смещением в сторону увеличения, поэтому *t*-статистка уменьшается и вывод о значимости коэффициентов ненадежен;
3. Оценки коэффициентов по МНК и стандартные ошибки становятся чувствительными к малейшим изменениям данных.
4. Затрудняется определение вклада каждой из объясняющих переменных в изменения эндогенной переменной.

Методы выявления мультиколлинеарности:

1. Если коэффициент детерминации R^2 достаточно высок, но некоторые из коэффициентов регрессии незначимы, т.е. они имеют низкие *t*-статистики, то можно говорить о присутствии мультиколлинеарности в модели *по формальным признакам*.

2. Для выявления мультиколлинеарности в регрессионных моделях возможно использовать коэффициенты корреляции: парные в случае парной линейной регрессии и частные в случае множественной линейной регрессии.

Парная корреляция между малозначимыми факторами x_i достаточно высока. Однако данное предположение будет выполняться лишь в случае двух объясняющих переменных. При большом их количестве надежнее использовать частные коэффициенты корреляции. Коэффициент корреляции определяется выражением:

$$corr(x;y) = \frac{\operatorname{cov}(x;y)}{\sigma_x \cdot \sigma_y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2} \cdot \sqrt{\overline{y^2} - \bar{y}^2}},$$

где: $\operatorname{cov}(x;y)$ – ковариация случайных величин x и y;

σ_x и σ_y – стандартные отклонения случайных величин x и y.

Коэффициент корреляции является безразмерной величиной и показывает степень линейной связи двух переменных: $r > 0$ при положительной связи и $r = 1$ при строгой положительной линейной связи; $r < 0$ при отрицательной связи и $r = -1$ при строгой отрицательной линейной связи; $r = 0$ при отсутствии линейной связи.

Случайные величины x и y называются ***некоррелированными***, если $r = 0$, и коррелированными, если $r \neq 0$. Независимые случайные величины x и y всегда некоррелированные (т.е. $r = 0$), но из некоррелированности случайных величин x и y не следует их независимость. Некоррелированность указывает лишь на отсутствие линейной связи между переменными, но не на отсутствие связи между ними вообще. Считается, что две переменные явно коллинеарны, если $corr(x_1;x_2) \geq 0{,}7$. В этом случае факторы дублируют друг друга, и один из них рекомендуется исключить из регрессии. Предпочтение при этом отдаётся фактору, который при достаточно тесной связи с результатом имеет наименьшую тесноту связи с другими факторами.

Частные коэффициенты корреляции характеризуют тесноту связи между результатом и соответствующим фактором при устранении влияния других факторов, включенных в уравнение регрессии. Показатели частной корреляции представляют собой отношение сокращения остаточной дисперсии за счет дополнительного включения в анализ нового фактора к остаточной дисперсии, имевшей место до введения его в модель. Например, для случая трех объясняющих переменных X_1, X_2, X_3 необходимо рассмотреть три частных коэффициента корреляции: $r_{12.3}$ – оценивает корреляцию между X_1 и X_2, исключая их связь с переменной X_3; $r_{13.2}$ –

оценивает корреляцию между X_1 и X_3, исключая их связь с переменной X_2; $r_{23.1}$ – оценивает корреляцию между X_2 и X_3, исключая их связь с переменной X_1. Найти значения частных коэффициентов корреляции можно с помощью найденных значений парных коэффициентов корреляции между переменными, обозначенных соответственно для удобства $r_{12} = corr(X_1; X_2)$, $r_{13} = corr(X_1; X_3)$ и $r_{23} = corr(X_2; X_3)$. Формулы для расчета частных коэффициентов корреляции имеют вид:

$$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{1 - r_{13}^2} \cdot \sqrt{1 - r_{23}^2}}; \; r_{13.2} = \frac{r_{13} - r_{12} \cdot r_{23}}{\sqrt{1 - r_{12}^2} \cdot \sqrt{1 - r_{23}^2}}; \; r_{23.1} = \frac{r_{23} - r_{12} \cdot r_{13}}{\sqrt{1 - r_{12}^2} \cdot \sqrt{1 - r_{13}^2}}.$$

При количестве экзогенных переменных $m > 3$ целесообразно применять методы матричного исчисления. Процедура нахождения частных коэффициентов корреляции проводится в несколько шагов:

1 шаг: находят корреляционную матрицу, т.е. матрицу парных коэффициентов корреляции (помним при этом, что матрица является симметричной, т.е. $r_{ij} = r_{ji}$):

$$r = \begin{pmatrix} 1 & r_{12} & r_{13} & \dots & r_{1m} \\ r_{21} & 1 & r_{23} & \dots & r_{2m} \\ r_{31} & r_{32} & 1 & \dots & r_{3m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & r_{m3} & \dots & 1 \end{pmatrix}$$

2 шаг: находят матрицу, обратную к корреляционной матрице $C = r^{-1}$:

$$C = r^{-1} = \begin{pmatrix} c_{11} & c_{12} & c_{13} & \dots & c_{1m} \\ c_{21} & c_{22} & c_{23} & \dots & c_{2m} \\ c_{31} & c_{32} & c_{33} & \dots & c_{3m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & c_{m3} & \dots & c_{mm} \end{pmatrix}$$

3 шаг: частные коэффициенты корреляции рассчитываются по формуле:

$r_{ij} = \dfrac{-c_{ij}}{\sqrt{c_{ii} \cdot c_{jj}}}$, где r_{ij} – частный коэффициент корреляции между переменными x_i и x_j, очищенный от их корреляции со всеми остальными экзогенными переменными.

Также коэффициенты частной корреляции более высоких порядков можно определить через коэффициенты частной корреляции более низких порядков по рекуррентным формулам (по аналогии с тем, как показано вычисление частных коэффициентов корреляции при $m = 3$ через коэффициенты парной корреляции).

3. Для выявления мультиколлинеарности возможно использовать собственные значения матрицы, обратной к матрице перекрестных

произведений экзогенных переменных, и так называемый индекс условности. В рамках данного подхода определяются собственные значения $\left(X^T X\right)^{-1}$, затем рассчитывается условное число $k = \frac{максимальное\,собственное\,число}{минимальное\,собственное\,число}$ и находится значение индекс аусловности $CI = \sqrt{k}$. В зависимости от получаемых значений k и CI делается вывод о наличии мультиколлинеарности:

$100 < k < 1000$ или $10 < CI < 30$ – умеренная мультиколлинеарность;

$k > 1000$ или $CI > 30$ – сильная мультиколлинеарность.

4. Наиболее часто при выявлении мультиколлинеарности используется метод вариации дисперсионно-инфляционного фактора. Для каждой экзогенной переменной x_j находится значение коэффициента вариации инфляционного фактора:

$$VIF_j = \frac{1}{1 - R_j^2},$$

где R_j^2 – коэффициент детерминации вспомогательной регрессии переменной x_j на все остальные экзогенные переменные. Значение коэффициента VIF_j сравнивается с критическим значением $VIF = 10$. Если $VIF_j < 10$, то связь между j-факторами и i-факторами недостаточно сильная для того, чтобы говорить о мультиколлинеарности; если $VIF_j \geq 10$, то в модели присутствует мультиколлинеарности, связанная с переменной x_j. Согласно некоторым источникам, к значениям коэффициента VIF_j можно применять более жесткие требования и сравнить его со значением $VIF = 5$. Отличие метода вариации инфляционного фактора от, например, использования частных коэффициентов корреляции в том, что метод непосредственно указывает на переменные, включение которых в модель приводит к мультиколлинеарности. При это необязательно строить вспомогательные модели регрессии, поскольку при вычислении матрицы С, обратной к корреляционной матрице, полученные элементы главной диагонали представляют собой соответствующие значения коэффициентов, т.е. $VIF_j = c_{jj}$.