$$Corr(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x) \cdot Var(}}$$

Эконометрическое моделирование

Лабораторная работа № 6

Анализ остатков. Гетероскедастичность

Оглавление

Свойства остатков

Свойства коэффициентов регрессии существенным образом зависят от свойств случайного члена. Для того, чтобы регрессионный анализ, основанный на обычном методе наименьших квадратов, давал наилучшие из всех возможных результаты, случайный член должен удовлетворять четырем условиям, известным как условия Гаусса-Маркова. Если эти условия не выполнены, исследователь должен это сознавать. Если корректирующие действия возможны, то аналитик должен быть в состоянии их выполнить. Если ситуацию исправить невозможно, исследователь должен быть способен судить, насколько серьезно это может повлиять на результаты. Рассмотрим теперь эти условия.

1-е условие Гаусса-Маркова: $E(\varepsilon_i) = 0$ для всех наблюдений

Первое условие состоит в том, что *математическое ожидание* случайного члена (остатков) в любом наблюдении должно быть равно нулю. Иногда случайный член будет положительным, иногда отрицательным, но он не должен иметь систематического смещения ни в каком из двух возможных направлений. Фактически, если уравнение регрессии включает постоянный член, то обычно бывает разумно предположить, что это условие выполняется автоматически, так как роль постоянного члена состоит в отражении любой систематической, но постоянной составляющей в Y, которую не учитывают объясняющие переменные, включенные в уравнение регрессии.

Для того чтобы проверить данную предпосылку достаточно найти математическое ожидание остатков и убедиться, что оно близко к 0.

Задание 1. Проверка выполнения 1-го условия Гаусса-Маркова

Проверьте, выполняется ли данное условие для модели, полученной вами в лабораторной работе №5.

2-е условие Гаусса-Маркова: теоретическая дисперсия ε_i, постоянна для всех наблюдений

Второе условие состоит в том, что дисперсия случайного члена должна быть постоянна для всех наблюдений. Иногда случайный член будет больше, иногда меньше, однако не должно быть априорной причины для того, чтобы он порождал большую ошибку в одних наблюдениях, чем в других. Такое утверждение может показаться странным и требует пояснение. Случайный член в каждом наблюдении имеет только одно значение, и может возникнуть вопрос о том, что означает его «дисперсия». Имеется в виду его возможное поведение до того, как сделана выборка. Когда мы записываем модель

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

первые два условия Гаусса-Маркова указывают, что случайные члены ε_i в n наблюдениях появляются на основе вероятностных распределений, имеющих нулевое математическое ожидание и одну и ту же дисперсию. Их фактические значения в выборке

иногда будут положительными, иногда – отрицательными, иногда – относительно далекими от нуля, иногда – относительно близкими к нему, но у нас нет причин заранее ожидать появления особенно больших отклонений в любом данном наблюдении. Другими словами, вероятность того, что величина ε примет какое-то данное положительное (или отрицательное) значение, будет одинаковой для всех наблюдений. Это условие известно *гомоскедастичность*, что означает «одинаковый разброс».

Рассмотрим несколько рисунков, которые иллюстрируют гомоскедастичность.

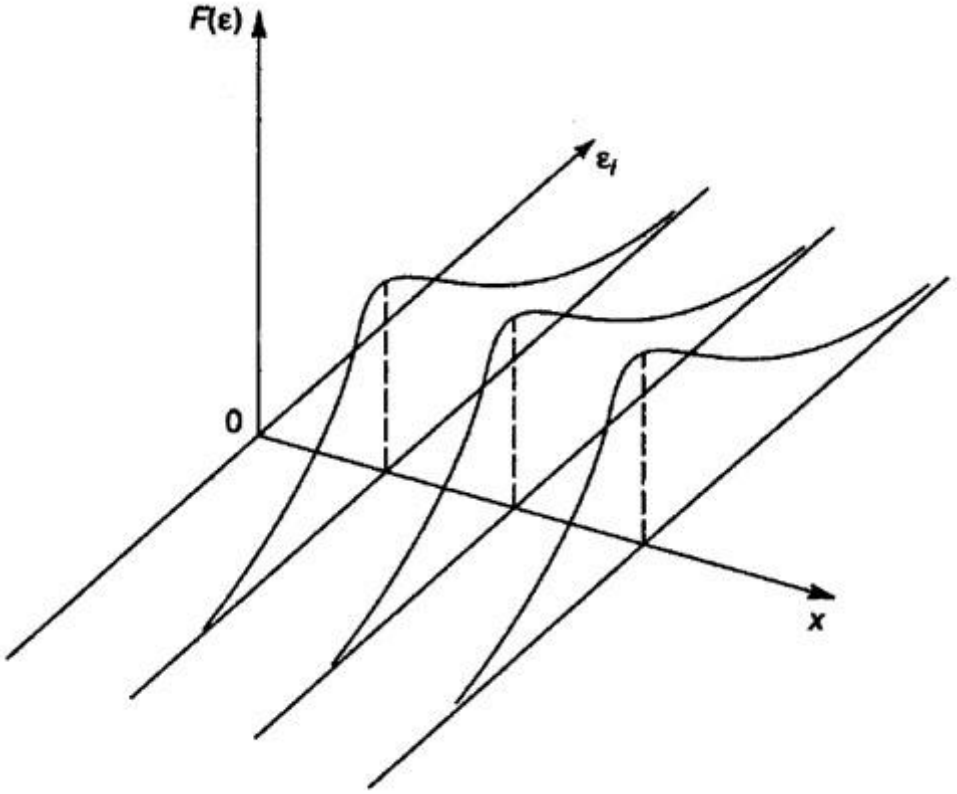

Рисунок 1 – Иллюстрация гомоскедастичности

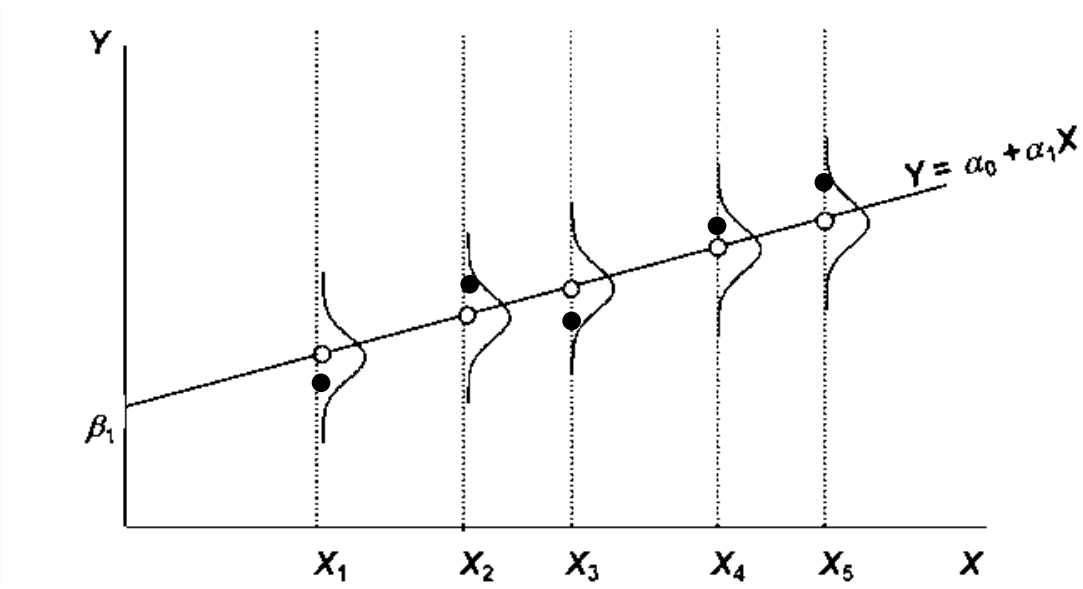

Рисунок 2 – Иллюстрация гомоскедастичности

Рассмотрим подробнее иллюстрацию гомоскедастичности, оказанную на рис. 2. Чтобы рисунок был достаточно простым, в выборку включено лишь пять наблюдений. Начнем с первого наблюдения, в котором переменная X принимает значение X_1. Потенциальное распределение случайного члена, определяющее формирование очередного наблюдения, представлено нормальным распределением с центром в соответствующем кружочке. Фактическое значение случайного члена в первом наблюдении оказалось отрицательным, и это наблюдение показано черным кружочком. Потенциальное распределение случайного члена и фактическое наблюдение представлены аналогичным образом и для остальных четырех наблюдений.

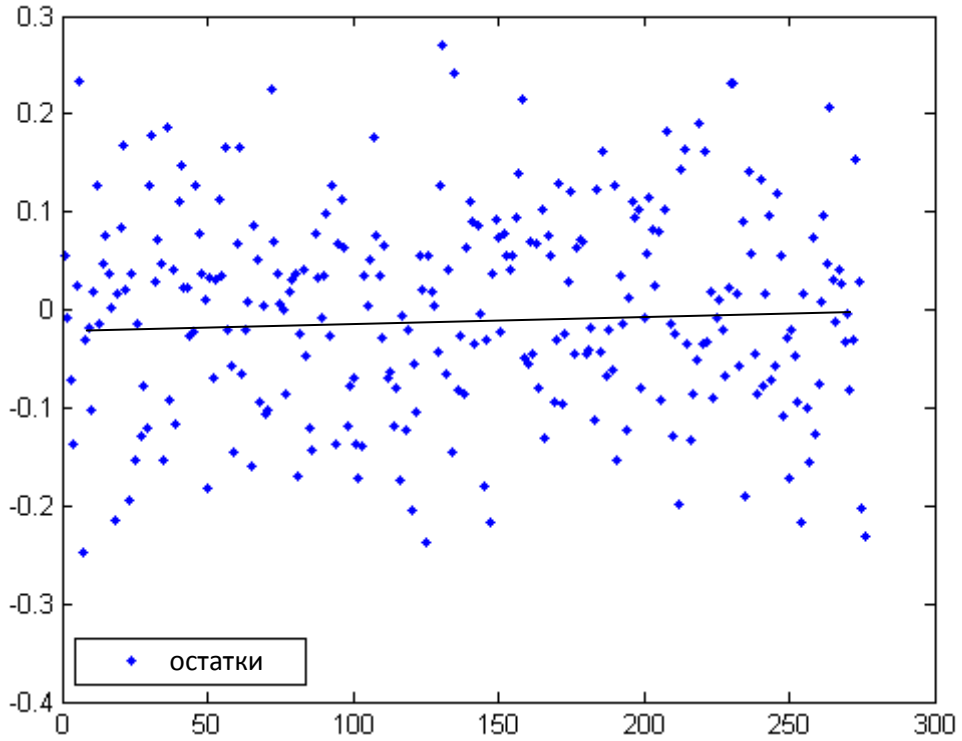

Рисунок 3 – Модель с гомоскедастичными остатками

Хотя гомоскедастичность в регрессионном анализе часто рассматривается как данная, в некоторых случаях более реалистичным оказывается предположение, что потенциальное распределение случайного члена в разных наблюдениях выборки различно. Это показано на рис. 4, где дисперсия потенциального распределения случайного члена возрастает по мере возрастания X. Это не означает, что случайный член обязательно будет иметь особенно большие (положительные или отрицательные) значения в тех наблюдениях, где значение X велико, но это значит, что априорная вероятность получения сильно отклоненных величин будет относительно высока. Это пример «гетероскедастичносми» что означает «неодинаковый разброс».

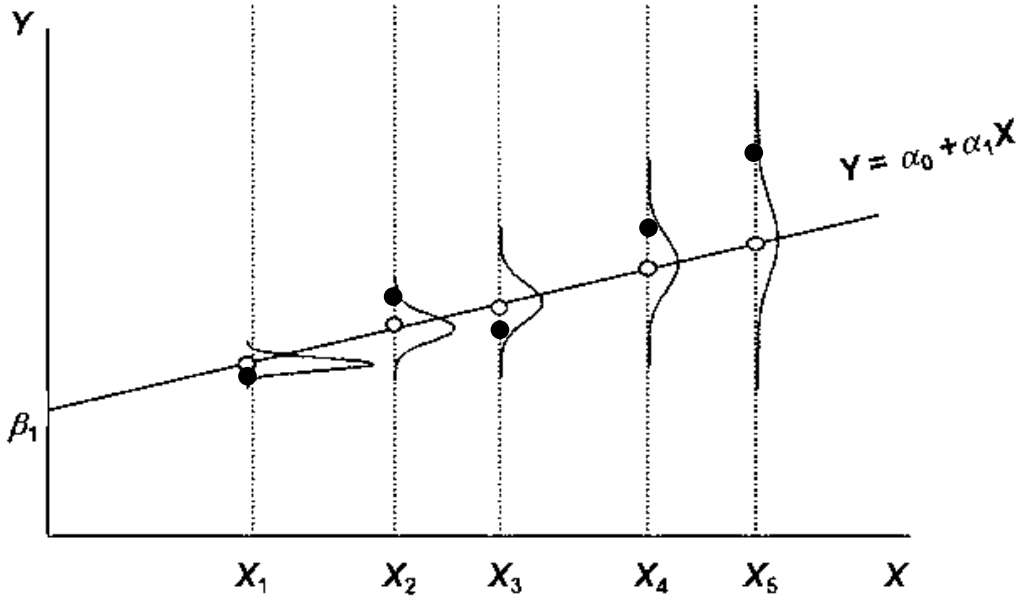

Рисунок 4 – Иллюстрация гетероскедастичности

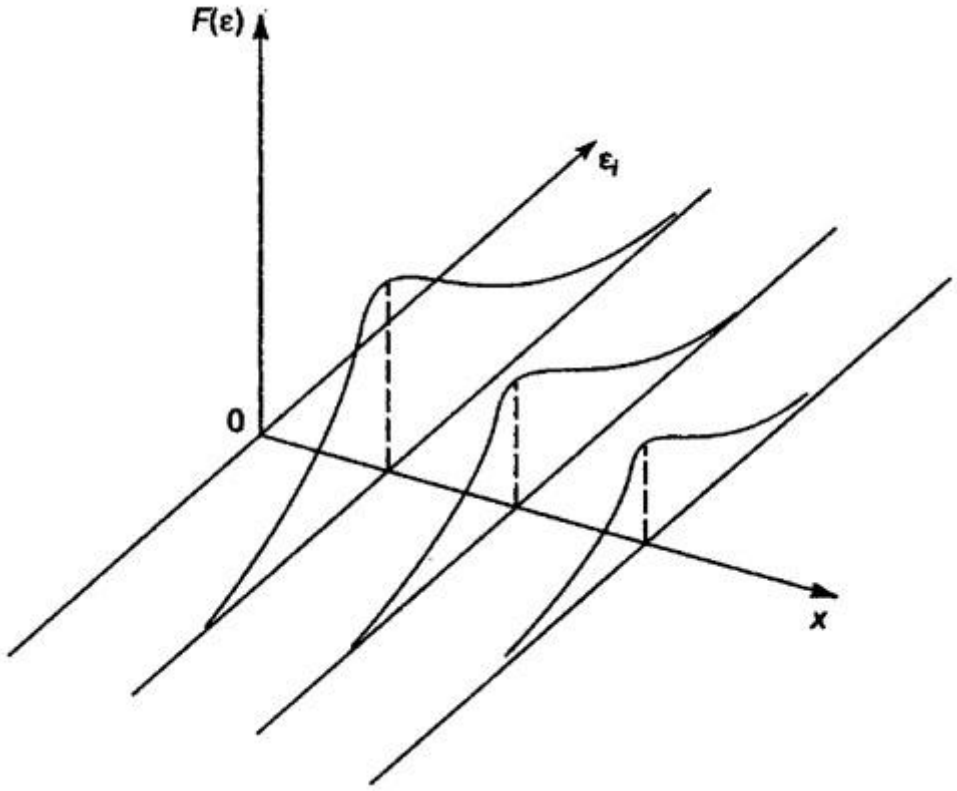

Рисунок 5 – Иллюстрация гетероскедастичности

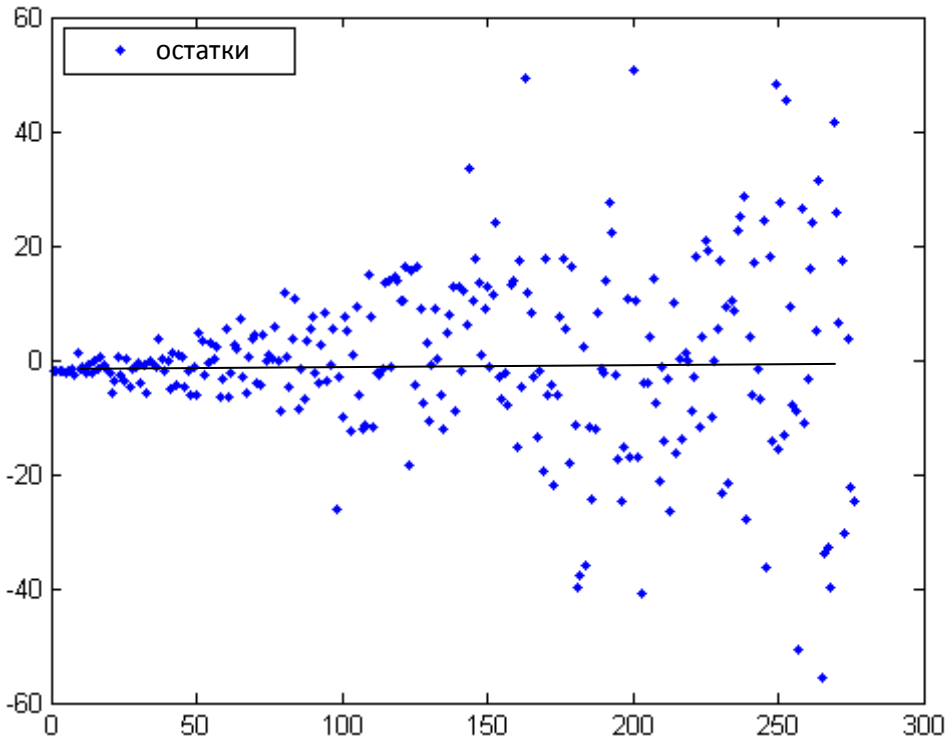

Рисунок 6 – Модель с гетероскедастичными остатками

Рассмотрим, почему гетероскедастичность имеет существенное значение. Первая причина касается дисперсий коэффициентов регрессии. Желательно, чтобы они были как можно меньше, т. е. (в вероятностном смысле) обеспечивали максимальную точность. При отсутствии гетероскедастичности и выполнении остальных условий Гаусса-Маркова полученные по МНК коэффициенты регрессии имеют наименьшую дисперсию среди всех несмещенных оценок. Если имеет место гетероскедастичность, то *МНК-оценки неэффективны*, поскольку можно найти другие оценки, которые имеют меньшую дисперсию и, тем не менее, являются несмещенными.

Вторая, не менее важная, причина заключается в том, что сделанные *оценки стандартных ошибок коэффициентов регрессии* будут *неверны*. Они вычисляются на основе предположения о том, что распределение случайного члена гомоскедастично. Если это не так, то они оказываются смещены, и вследствие этого *t*-критерии и обычный F-критерий неприменимы. Вполне вероятно, что стандартные ошибки будут занижены, а следовательно, t-статистика – завышена, и будет получено неправильное представление о точности коэффициентов регрессии.

Обнаружение гетероскедастичности

Рассмотрим три обычно используемых теста (критерия), в которых делаются различные предположения о зависимости между дисперсией случайного члена и величиной объясняющей переменной (переменных): тест ранговой корреляции Спирмена, тест Голдфелда-Квандта и тест Глейзера.

Тест ранговой корреляции Спирмена

При выполнении теста ранговой корреляции Спирмена предполагается» что дисперсия случайного члена будет либо увеличиваться, либо уменьшаться по мере увеличения X, и поэтому в регрессии, оцениваемой с помощью МНК, абсолютные величины остатков и значения X будут коррелированны. Данные по X и абсолютные величины остатков упорядочиваются, и коэффициент ранговой корреляции определяется как

$$r_{Xe} = 1 - \frac{6\sum D_i^2}{n(n^2-1)}$$

где D_i – разность между рангом X и ε в наблюдении i.

Если предположить, что соответствующий коэффициент корреляции для генеральной совокупности равен нулю, то коэффициент ранговой корреляции имеет нормальное распределение с математическим ожиданием 0 и дисперсией 1/(n – 1) в больших выборках. Следовательно, соответствующая тестовая статистика равна $r_{Xe}\sqrt{n-1}$, и при использовании двустороннего критерия нулевая гипотеза о гомоскедастичности будет отклонена при уровне значимости 5%, если ее абсолютная величина превысит 1,96, и при уровне значимости 1%, если ее абсолютная величина превысит 2,58. Если в модели регрессии имеется более одной объясняющей переменной, то проверка гипотезы может выполняться с использованием любой из них.

Пример

Допустим, с помощью МНК оценена следующая регрессионная зависимость (в скобках приведены стандартные ошибки):

$$\hat{Y} = 604 + 0{,}194X$$

(5700) (0,013) R^2= 0,89

Проранжированные отклонения от линии регрессии и значения X приведены в табл.1, и на их основе вычислены показатели D_i и D_i^2.

Таблица 1

X	***Ранг***	***|ε|***	***Ранг***	***D***	***D²***	***X***	***Ранг***	***|ε|***	***Ранг***	**D**	**D²**
13746	1	547	2	-1	1	130 823	15	14185	23	-8	64
14386	2	1130	4	-2	4	135 961	16	4176	12	4	16
24 646	3	2620	8	-5	25	151 266	17	3976	11	6	36
41 506	4	1417	5	-1	1	198 432	18	4233	14	4	16
44 753	5	5955	15	-10	100	232 006	19	1025	3	16	256
50919	6	2629	9	-3	9	261 386	20	6270	17	3	9
52 662	7	6768	19	-12	144	334 286	21	16 758	24	-3	9
71 039	8	6264	18	-10	100	380 820	22	86 952	28	-6	36
72 60S	9	4227	13	-4	16	420 788	23	27 034	25	-2	4
74121	10	3611	10	0	0	483 652	24	14 180	22	2	4
87 352	11	499	1	10	100	547 203	25	6024	16	9	81
97 624	12	2067	6	6	36	1 016 286	26	62 439	27	-1	1

98 861	13	10 360	20	-7	49	1 024 609	27	45 333	26	1	1
122 926	14	10 929	21	-7	49	1 330 998	28	2093	7	21	441

Сумма D_i^2 составила 1608. Таким образом, коэффициент ранговой корреляции составляет

$$r_{Xe} = 1 - \frac{6 \cdot 1608}{28 \cdot 783} = 0{,}56$$

И тестовая статистика равно $0{,}56\sqrt{27} = 2{,}91$. Это превышает 2,58, и, следовательно, нулевая гипотеза о гомоскедастичности отвергается при уровне значимости 1%.

Тест Голдфелда-Квандта

Вероятно, наиболее популярным формальным критерием является критерий, предложенный С. Голдфелдом и Р. Квандтом (Goldfeld, Quandt, 1965). При проведении проверки по этому критерию предполагается, что стандартное отклонение ($\sigma_{\varepsilon i}$) распределения вероятностей случайного члена в наблюдении *i* пропорционально значению X_i. Предполагается также, что случайный член нормально распределен и удовлетворяет другим условиям Гаусса—Маркова.

Все *n* наблюдений в выборке упорядочиваются по величине *X*, после чего оцениваются отдельные регрессии для первых *n'* и для последних *n'* наблюдений; средние (*n* – *2n'*) наблюдений отбрасываются. Если имеет место гетероскедастичность и если предположение относительно ее природы верно, то дисперсия ε в последних *n'* наблюдениях будет больше, чем в первых *n'*, и это будет отражено в сумме квадратов остатков в двух указанных «частных» регрессиях. Обозначая суммы квадратов остатков в регрессиях для первых *n'* и последних *n'* наблюдений соответственно через RSS_1 и RSS_2, рассчитаем отношение RSS_2/RSS_1 которое при выполнении нулевой гипотезы о гомоскедастичности имеет *F*-распределение с (*n'* – *k*) и (*n'* – *k*) степенями свободы, где *k* – число объясняющих переменных в регрессионном уравнении. Мощность критерия зависит от выбора *n'* по отношению к *n*. Голдфелд и Квандт, основываясь на результатах некоторых проведенных ими экспериментов, утверждают, что *n'* должно составлять порядка 3/8 от *n* частности около 11, если *n* = 30, и около 22, если *n* = 60. Если в модели имеется более одной объясняющей переменной, то наблюдения должны упорядочиваться по той из них, которая, как предполагается, связана с $\sigma_{\varepsilon i}$.

Нулевая гипотеза для данного теста состоит в том, что RSS_2 не превышает значимо RSS_1 а альтернативная гипотеза – значимо превышает. Если величина RSS_2 оказалась меньшей, чем RSS_1 то вы не можете отвергнуть нулевую гипотезу, и вычислять тестовую статистику RSS_2/RSS_1 нет нужды. Однако метод Голдфелда-Квандта может также использоваться для проверки на гетероскедастичность при предположении, что стандартное отклонение случайного члена обратно пропорционально X_i. При этом используется та же процедура, что и описанная выше, но тестовой статистикой теперь является показатель RSS_1/RSS_2, который вновь имеет F-распределение с (*n'* – *k*) И (*n'* – *k*) степенями свободы при выполнении нулевой гипотезы о гомоскедастичности.

Пример.

На основе данных табл. 2 с помощью обычного МНК были оценены регрессии сначала по наблюдениям для 11 стран с наименьшим X, а затем – для 11 стран с наибольшим X. Сумма квадратов отклонений в первой регрессии была равна 157 x 10^6, а во второй – 13,518 x 10^6. Отношение RSS_2/RSS_1, следовательно, составило 86,1. Критическое значение *F(9;* 9) при уровне значимости 0,1% составляет 10,1, и следовательно, нулевая гипотеза о гомоскедастичности была отклонена.

Таблица 2

Страна	Y	X
Бельгия	44 517	232 006
Канада	112617	547 203
Чили	13 096	50 919
Дания	25 927	151 266
Финляндия	21 581	97 624
Франция	256 316	1 330 998
Греция	9392	98 861
Гонконг	11 758	130 823
Венгрия	7227	41 506
Ирландия	17 572	52 662
Израиль	11 349	74 121
Италия	145013	1 016 286
Южная Корея	161 318	380 820
Кувейт	2797	24 848
Малайзия	18 874	72 505
Мексика	55 073	420 788
Нидерланды	48 595	334 286
Норвегия	13 484	122 926
Португалия	17 025	87 352
Сингапур	20 648	71 039
Словакия	2720	13 746
Словения	4520	14 386
Испания	80 104	483 652
Швеция	34 806	198 432
Швейцария	57 503	261 388
Сирия	3317	44 753
Турция	31 115	135 961
Великобритания	244 397	1 024 609

Тест Глейзера

Тест Глейзера позволяет несколько более тщательно рассмотреть характер гетероскедастичности. Мы снимаем предположение о том, что $\sigma_{\varepsilon i}$ пропорционально X_i и хотим проверить, может ли быть более подходящей какая-либо другая функциональная форма, например

$$\varepsilon_i = \beta_0 + \beta_1 X_i^{\gamma} \qquad (1)$$

Чтобы использовать данный метод, следует оценить регрессионную зависимость *Y* от *X* с помощью обычного МНК, а затем оценить регрессию *абсолютных* величин остатков

|ε|, оценив их регрессию вида (1) для данного значения Υ. Можно оценить несколько таких уравнений регрессии, изменяя значение Υ. В каждом случае нулевая гипотеза о гомоскедастичности будет отклонена, если оценка β_1 значимо отличается от нуля. Если при оценивании более чем одной функции получается значимая оценка β_1, то ориентиром при определении характера гетероскедастичности может служить наилучшая из них.

Пример

На основе данных по H и |ε| из табл. 1 были оценены уравнения (1) с использованием значений γ от -1,0 до 1,5. Результаты представлены в обобщенном виде в табл. 3. Следует отметить, что различные оценки β_1 несравнимы, так как определение объясняющей переменной X^{Υ}, в каждом случае разное. Тем не менее, уровни R^2 здесь сравнимы в том смысле, что зависимая переменная в каждом случае одна и та же. Статистически значимые на уровне 1% оценки были получены для трех средних значений Υ. Наилучшие результаты соответствуют значениям Υ, равным 0,25 и 0,5, и, следовательно, стандартное отклонение распределения ε возрастает с увеличением ВВП, но не в той же пропорции.

Таблица 3

Υ	β_1	*с.о.(β_1)*	R^2
-1,0	$-3{,}51 \times 10^{8}$	$1{,}94 \times 10^{8}$	0,11
-0,5	$-4{,}21 \times 10^{6}$	$1{,}71 \times 10^{6}$	0,19
-0,25	$0{,}56 \times 10^{6}$	0,20 хЮ^{6}	0,23
0,25	**1640**	**520**	**0,28**
0,5	**36,1**	**11.8**	**0,27**
1,0	0,026	0,010	0,21
1,5	$19{,}8 \times 10^{-6}$	$9{,}3 \times 10^{-6}$	0,15

Задание 2. Проверка выполнения 2-го условия Гаусса-Маркова

Проверьте, выполняется ли данное условие для модели, полученной вами в лабораторной работе №5. Используйте все три вида тестов.

Устранение гетероскедастичности

Для устранения гетероскедастичности можно применить масштабирование переменных или использовать нелинейные формы взаимосвязи. Данные способы подробно рассмотрены в учебнике Доугерти К. Введение в эконометрику. М: ИНФРА-М, 2007, С.235–241.

Способ устранения гетероскедастичности с помощью масштабирования переменных носит название «Взвешенный метод наименьших квадратов». Рассмотрите, как данные метод описан в презентации в сети Интернет.

Следующие две предпосылки относительно остатков регрессии будут рассмотрены в лабораторной работе №7.