

8. ГЕТЕРОСКЕДАСТИЧНОСТЬ

При проведении регрессионного анализа, основанного на методе наименьших квадратов, на практике следует обратить серьезное внимание на проблемы, связанные с выполнимостью свойств случайных отклонений моделей. Как мы отмечали ранее, свойства оценок коэффициентов регрессии напрямую зависят от свойств случайного члена в уравнении регрессии. Для получения качественных оценок необходимо следить за выполнимостью предпосылок МНК (условий Гаусса–Маркова), т. к. при их нарушении МНК может давать оценки с плохими статистическими свойствами. При этом существуют другие методы определения более точных оценок. Одной из ключевых предпосылок МНК является условие постоянства дисперсий случайных отклонений (см. параграф 5.1, предпосылка 2⁰):

дисперсия случайных отклонений ε_i постоянна. $D(\varepsilon_i) = D(\varepsilon_j) = \sigma^2$ для любых наблюдений i и j .

Выполнимость данной предпосылки называется *гомоскедастичностью (постоянством дисперсии отклонений)*. Невыполнимость данной предпосылки называется *гетероскедастичностью (непостоянством дисперсий отклонений)*.

В данной главе мы подробно проанализируем суть гетероскедастичности, ее причины и последствия, а также приведем несколько способов смягчения этих последствий.

8.1. Суть гетероскедастичности

При рассмотрении выборочных данных требование постоянства дисперсии случайных отклонений может вызвать определенное недоумение в силу того, что при каждом i -м наблюдении имеется единственное значение ε_i . Откуда же появляется разброс? Дело в том, что при рассмотрении выборочных данных мы имеем дело с конкретными реализациями зависимой переменной y_i и соответственно с определенными случайными отклонениями ε_i , $i = 1, 2, \dots, n$. Но до осуществления выборки эти показатели априори могли принимать произвольные значения на основе некоторых вероятностных распределений. Одним из требований к этим распределениям является равенство дисперсий. Данное условие подразумевает, что несмотря на то что при каждом конкретном наблюдении случайное отклонение может быть большим либо маленьким, положительным либо отрицательным, не должно быть некой априорной причины, вызывающей большую

ошибку (отклонение) при одних наблюдениях и меньшую – при других.

Однако на практике гетероскедастичность не так уж и редка. Зачастую есть основания считать, что вероятностные распределения случайных отклонений ε_i при различных наблюдениях будут различными. Это не означает, что случайные отклонения обязательно будут большими при определенных наблюдениях и малыми – при других, но это означает, что априорная вероятность этого велика. Поэтому важно понимать суть этого явления и его последствия.

На рис. 8.1 приведены два примера линейной регрессии – зависимости потребления C от дохода I : $C = \beta_0 + \beta_1 I + \varepsilon$.

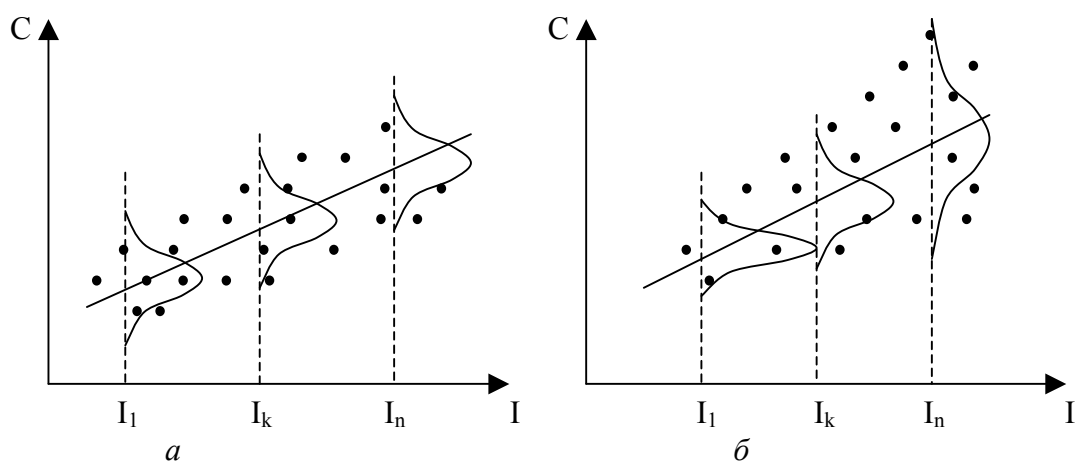


Рис. 8.1

В обоих случаях с ростом дохода растет среднее значение потребления. Но если на рис. 8.1, *а* дисперсия потребления остается одной и той же для различных уровней дохода, то на рис. 8.1, *б* при аналогичной зависимости среднего потребления от дохода дисперсия потребления не остается постоянной, а увеличивается с ростом дохода. Фактически это означает, что во втором случае субъекты с большим доходом в среднем потребляют больше, чем субъекты с меньшим доходом, и, кроме того, разброс в их потреблении более существенен для большего уровня дохода. Фактически люди с большими доходами имеют больший простор для распределения своего дохода. Реальность данной ситуации не вызывает сомнений. Разброс значений потребления вызывает разброс точек наблюдения относительно линии регрессии, что и определяет дисперсию случайных отклонений. Динамика изменения дисперсий (распределений) отклонений для данного примера проиллюстрирована на рис. 8.2. При гомоскедастичности

(рис. 8.2, *а*) дисперсии ε_i постоянны, а при гетероскедастичности (рис. 8.2, *б*) дисперсии ε_i изменяются (в нашем примере – увеличиваются).

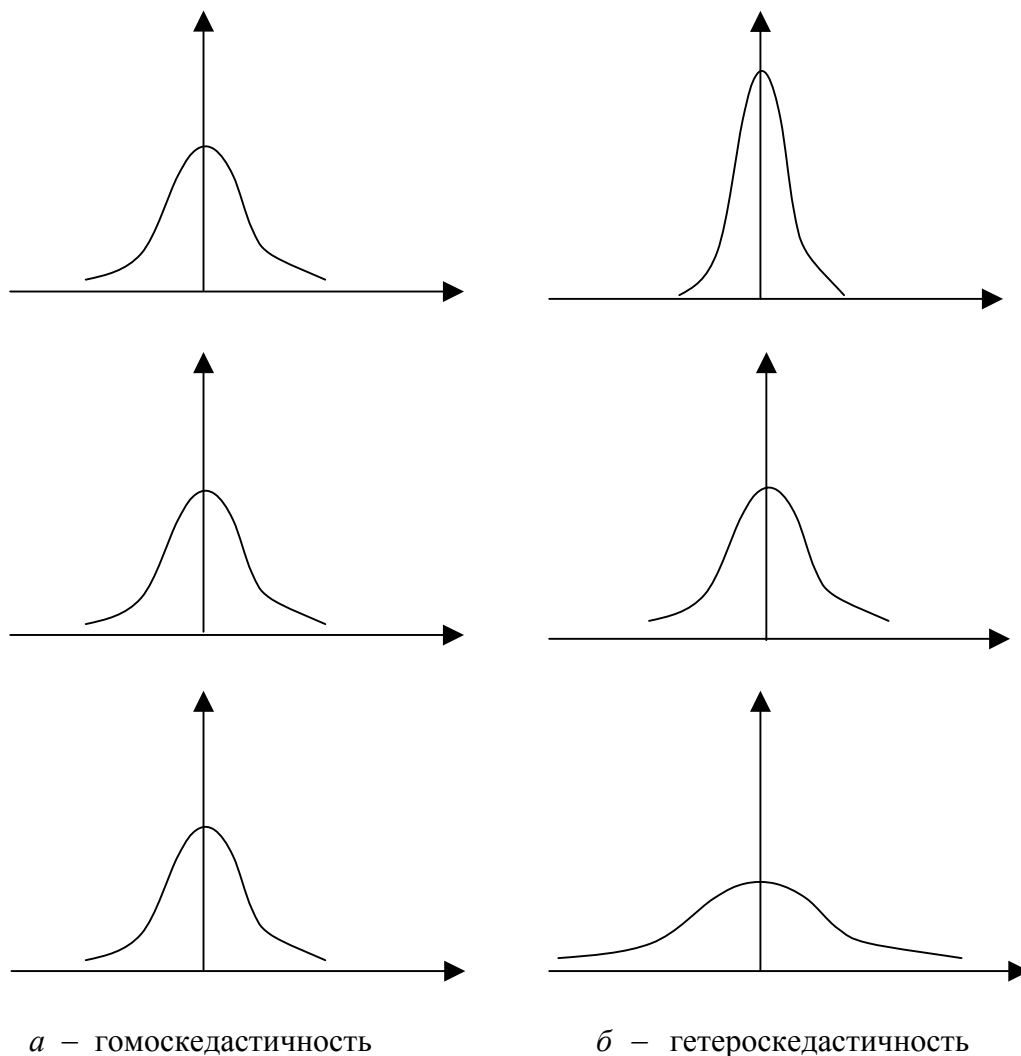


Рис. 8.2

Проблема гетероскедастичности в большей степени характерна для перекрестных данных и довольно редко встречается при рассмотрении временных рядов. Это можно объяснить следующим образом. При перекрестных данных учитываются экономические субъекты (потребители, домохозяйства, фирмы, отрасли, страны и т. п.), имеющие различные доходы, размеры, потребности и т. д. Но в этом случае возможны проблемы, связанные с эффектом масштаба. Во временных рядах обычно рассматриваются одни и те же показатели в различные моменты времени (например, ВВП, чистый экспорт, темпы инфляции

и т. д. в определенном регионе за определенный период времени). Однако при увеличении (уменьшении) рассматриваемых показателей с течением времени может возникнуть проблема гетероскедастичности.

8.2. Последствия гетероскедастичности

Как отмечалось в разделе 5.1, при рассмотрении классической линейной регрессионной модели МНК дает наилучшие линейные несмещенные оценки (BLUE-оценки) лишь при выполнении ряда предпосылок, одной из которых является постоянство дисперсии отклонений (гомоскедастичность): $\sigma^2(\varepsilon_i) = \sigma^2$ для всех наблюдений i , $i = 1, 2, \dots, n$.

При невыполнимости данной предпосылки (при гетероскедастичности) последствия применения МНК будут следующими.

1. Оценки коэффициентов по-прежнему остаются несмещенными и линейными.
2. Оценки не будут эффективными (т. е. они не будут иметь наименьшую дисперсию по сравнению с другими оценками данного параметра). Они не будут даже асимптотически эффективными. Увеличение дисперсии оценок снижает вероятность получения максимально точных оценок.
3. Дисперсии оценок будут рассчитываться со смещением. Смещенность появляется вследствие того, что необъясненная уравнением регрессии дисперсия $S^2 = \frac{\sum e_i^2}{n - m - 1}$ (m – число объясняющих переменных), которая используется при вычислении оценок дисперсий всех коэффициентов (см. параграф 6.2, (6.23)), не является более несмещенной.
4. Вследствие вышесказанного все выводы, получаемые на основе соответствующих t - и F -статистик, а также интервальные оценки будут ненадежными. Следовательно, статистические выводы, получаемые при стандартных проверках качества оценок, могут быть ошибочными и приводить к неверным заключениям по построенной модели. Вполне вероятно, что стандартные ошибки коэффициентов будут занижены, а следовательно, t -статистики будут завышены. Это может привести к признанию статистически значимыми коэффициентов, таковыми на самом деле не являющимися.

Причину неэффективности оценок МНК при гетероскедастичности легко пояснить следующим примером парной регрессии.

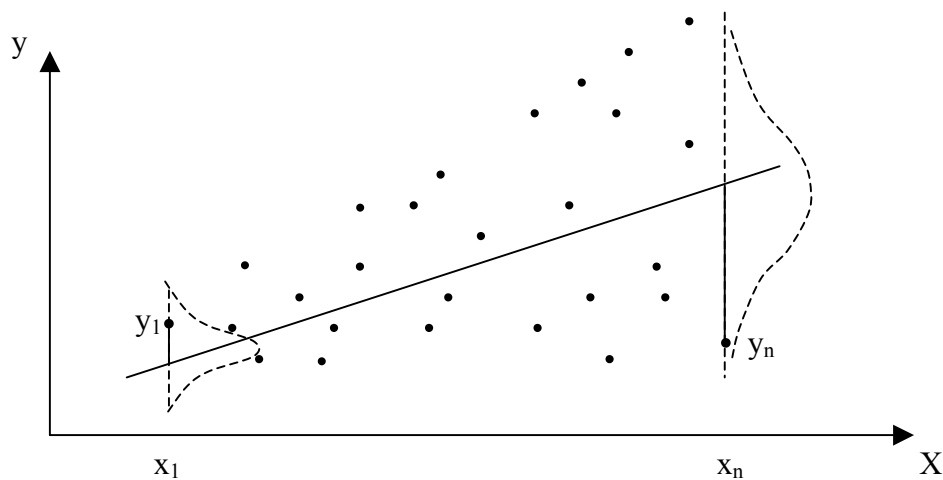


Рис. 8.3

Из рис. 8.3 видно, что для каждого конкретного значения x_i СВ X переменная Y принимает значение y_i из некоторого множества, имеющего свое распределение, отличное одно от другого в силу непостоянства дисперсий (сравните распределения для значений y_1 и y_n).

По МНК минимизируется сумма квадратов отклонений

$$\sum e_i^2 = \sum (y_i - b_0 - b_1 x_i)^2.$$

Но в этом случае каждое конкретное значение e_i^2 в данной сумме имеет одинаковый “вес” вне зависимости от того, получено оно из распределения с маленькой дисперсией (например, e_1^2) или с большой (например, e_n^2). Но это противоречит логике, т. к. точка, полученная из распределения с меньшей дисперсией, более точно определяет направление линии регрессии. Поэтому она должна иметь больший “вес”, чем точка из распределения с большей дисперсией. Следовательно, методы оценивания, учитывающие “веса” точек наблюдений, позволяют получать более точные (эффективные) оценки. Учет “весов” точек характерен, например, для метода взвешенных наименьших квадратов, рассмотренного ниже.

8.3. Обнаружение гетероскедастичности

В ряде случаев на базе знаний характера данных появление проблемы гетероскедастичности можно предвидеть и попытаться устранить этот недостаток еще на этапе спецификации. Однако значительно чаще эту проблему приходится решать после построения уравнения регрессии.

Обнаружение гетероскедастичности в каждом конкретном случае является довольно сложной задачей, т. к. для знания дисперсий отклонений $\sigma^2(e_i)$ необходимо знать распределение СВ Y , соответствующее выбранному значению x_i СВ X . На практике зачастую для каждого конкретного значения x_i определяется единственное значение y_i , что не позволяет оценить дисперсию СВ Y для данного x_i .

Естественно, не существует какого-либо однозначного метода определения гетероскедастичности. Однако к настоящему времени для такой проверки разработано довольно большое число тестов и критериев для них. Рассмотрим наиболее популярные и наглядные: графический анализ отклонений, тест ранговой корреляции Спирмена, тест Парка, тест Глейзера, тест Голдфелда–Квандта.

8.3.1. Графический анализ остатков

Использование графического представления отклонений позволяет определиться с наличием гетероскедастичности. В этом случае по оси абсцисс откладывается объясняющая переменная X (либо линейная комбинация объясняющих переменных $Y = b_0 + b_1X_1 + \dots + b_mX_m$), а по оси ординат либо отклонения e_i , либо их квадраты e_i^2 . Примеры таких графиков приведены на рис. 8.4.

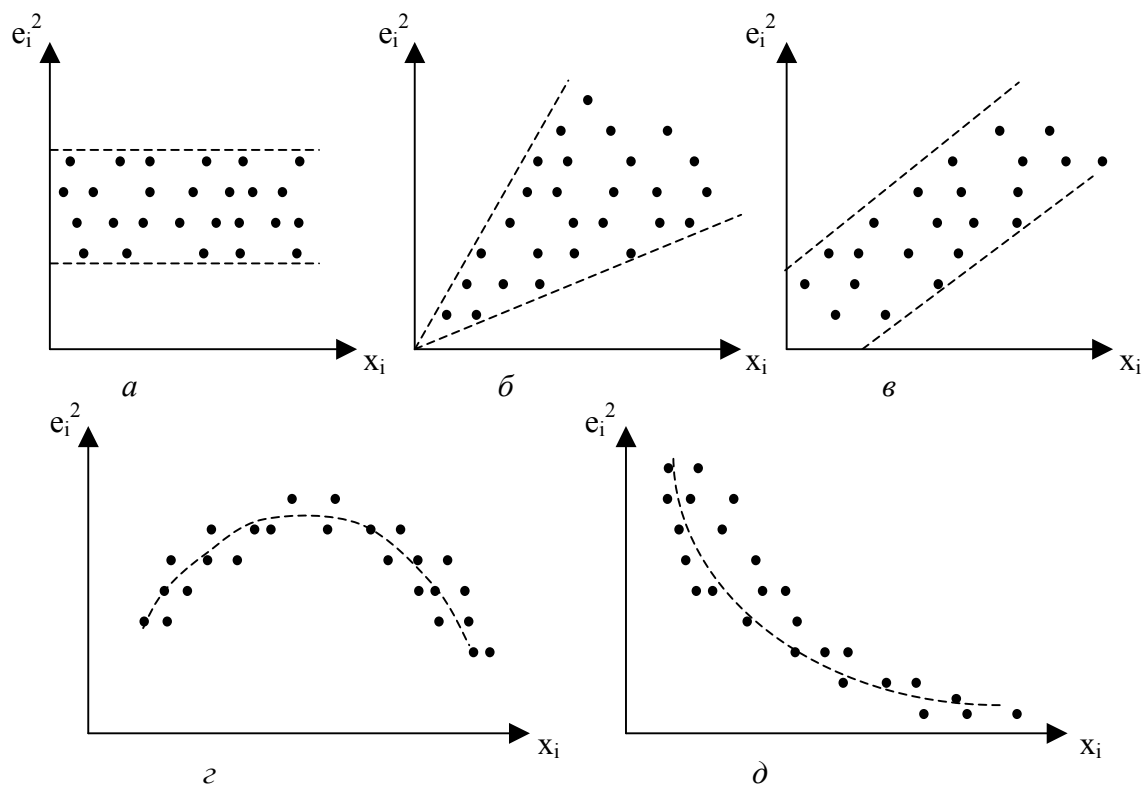


Рис. 8.4

На рис. 8.4, а все отклонения e_i^2 находятся внутри полуполосы постоянной ширины, параллельной оси абсцисс. Это говорит о независимости дисперсий e_i^2 от значений переменной X и их постоянстве, т.е. в этом случае мы находимся в условиях гомоскедастичности.

На рис. 8.4, б – г наблюдаются некие систематические изменения в соотношениях между значениями x_i переменной X и квадратами отклонений e_i^2 . Рис. 8.4, б соответствует примеру из параграфа 8.1. На рис. 8.4, в отражена линейная; 8.4, г – квадратичная; 8.4, д – гиперболическая зависимости между квадратами отклонений и значениями объясняющей переменной X . Другими словами, ситуации, представленные на рис. 8.4, б – д, отражают большую вероятность наличия гетероскедастичности для рассматриваемых статистических данных.

Отметим, что графический анализ отклонений является удобным и достаточно надежным в случае парной регрессии. При множественной регрессии графический анализ возможен для каждой из объясняющих переменных X_j , $j = 1, 2, \dots, m$ отдельно. Чаще же вместо объясняющих переменных X_j по оси абсцисс откладывают значения \hat{y}_i , получаемые из эмпирического уравнения регрессии. Поскольку по уравнению множественной линейной регрессии \hat{y}_i является линейной комбинацией x_{ij} , $j = 1, 2, \dots, m$, то график, отражающий зависимость e_i^2 от \hat{y}_i , может указать на наличие гетероскедастичности аналогично ситуациям на рис. 8.4, б – д. Такой анализ наиболее целесообразен при большом количестве объясняющих переменных.

8.3.2. Тест ранговой корреляции Спирмена

При использовании данного теста предполагается, что дисперсия отклонения будет либо увеличиваться, либо уменьшаться с увеличением значения X . Поэтому для регрессии, построенной по МНК, абсолютные величины отклонений e_i и значения x_i СВ X будут коррелированы. Значения x_i и e_i ранжируются (упорядочиваются по величинам). Затем определяется коэффициент ранговой корреляции:

$$r_{x,e} = 1 - 6 \cdot \frac{\sum d_i^2}{n(n^2 - 1)}, \quad (8.1)$$

где d_i – разность между рангами x_i и e_i , $i = 1, 2, \dots, n$; n – число наблюдений.

Например, если x_{20} является 25-м по величине среди всех наблюдений X ; а e_{20} – является 32-м, то $d_i = 25 - 32 = -7$.

Доказано, что если коэффициент корреляции $\rho_{x,e}$ для генеральной совокупности равен нулю, то статистика

$$t = \frac{r_{x,e} \sqrt{n-2}}{\sqrt{1-r_{x,e}^2}} \quad (8.2)$$

имеет распределение Стьюдента с числом степеней свободы $\nu = n - 2$.

Следовательно, если наблюдаемое значение t -статистики, вычисленное по формуле (8.2), превышает $t_{кр.} = t_{\alpha, n-2}$ (определяемое по таблице критических точек распределения Стьюдента), то необходимо отклонить гипотезу о равенстве нулю коэффициента корреляции $\rho_{x,e}$, а следовательно, и об отсутствии гетероскедастичности. В противном случае гипотеза об отсутствии гетероскедастичности принимается.

Если в модели регрессии больше чем одна объясняющая переменная, то проверка гипотезы может осуществляться с помощью t -статистики для каждой из них отдельно.

8.3.3. Тест Парка

Р. Парк предложил критерий определения гетероскедастичности, дополняющий графический метод некоторыми формальными зависимостями. Предполагается, что дисперсия $\sigma_i^2 = \sigma^2(e_i)$ является функцией i -го значения x_i объясняющей переменной. Парк предложил следующую функциональную зависимость

$$y_i^2 = y^2 x_i^b e^{v_i}. \quad (8.3)$$

Прологарифмировав (8.4), получим:

$$\ln y_i^2 = \ln y^2 + b \ln x_i + v_i. \quad (8.4)$$

Так как дисперсии y_i^2 обычно неизвестны, то их заменяют оценками квадратов отклонений e_i^2 .

Критерий Парка включает следующие этапы:

1. Строится уравнение регрессии $y_i = b_0 + b_1 x_i + e_i$.
2. Для каждого наблюдения определяются $\ln e_i^2 = \ln(y_i - \hat{y}_i)^2$.
3. Строится регрессия

$$\ln e_i^2 = \alpha + \beta \ln x_i + v_i, \quad (8.5)$$

где $\alpha = \ln \sigma^2$.

В случае множественной регрессии зависимость (8.5) строится для каждой объясняющей переменной.

4. Проверяется статистическая значимость коэффициента β уравнения (8.5) на основе t-статистики $t = \frac{\beta}{S_{\beta}}$. Если коэффициент β статистически значим, то это означает наличие связи между $\ln e_i^2$ и $\ln x_i$, т. е. гетероскедастичности в статистических данных.

Отметим, что использование в критерии Парка конкретной функциональной зависимости (8.5) может привести к необоснованным выводам (например, коэффициент β статистически незначим, а гетероскедастичность имеет место). Возможна еще одна проблема. Для случайного отклонения v_i в свою очередь может иметь место гетероскедастичность. Поэтому критерий Парка дополняется другими тестами.

8.3.4. Тест Глейзера

Тест Глейзера по своей сути аналогичен тесту Парка и дополняет его анализом других (возможно, более подходящих) зависимостей между дисперсиями отклонений σ_i и значениями переменной x_i . По данному методу оценивается регрессионная зависимость модулей отклонений $|e_i|$ (тесно связанных с σ_i^2) от x_i . При этом рассматриваемая зависимость моделируется следующим уравнением регрессии:

$$|e_i| = \alpha + \beta x_i^k + v_i. \quad (8.6)$$

Изменяя значения k , можно построить различные регрессии. Обычно $k = \dots, -1, -0.5, 0.5, 1, \dots$. Статистическая значимость коэффициента β в каждом конкретном случае фактически означает наличие гетероскедастичности. Если для нескольких регрессий (8.6) коэффициент β оказывается статистически значимым, то при определении характера зависимости обычно ориентируются на лучшую из них.

Отметим, что так же, как и в тесте Парка, в тесте Глейзера для отклонений v_i может нарушаться условие гомоскедастичности. Однако во многих случаях предложенные модели являются достаточно хорошими для определения гетероскедастичности.

8.3.5. Тест Голдфелда–Квандта

В данном случае также предполагается, что стандартное отклонение $\sigma_i = \sigma(\varepsilon_i)$ пропорционально значению x_i переменной X в этом наблюдении, т. е. $y_i^2 = y^2 x_i^2$. Предполагается, что ε_i имеет нормальное распределение и отсутствует автокорреляция остатков.

Тест Голдфелда–Квандта состоит в следующем:

1. Все n наблюдений упорядочиваются по величине X .
2. Вся упорядоченная выборка после этого разбивается на три подвыборки размерностей k , $(n - 2k)$, k соответственно.
3. Оцениваются отдельные регрессии для первой подвыборки (k первых наблюдений) и для третьей подвыборки (k последних наблюдений). Если предположение о пропорциональности дисперсий отклонений значениям X верно, то дисперсия регрессии (сумма квадратов отклонений $S_1 = \sum_{i=1}^k e_i^2$) по первой подвыборке будет существенно меньше дисперсии регрессии (суммы квадратов отклонений $S_3 = \sum_{i=n-k}^n e_i^2$) по третьей подвыборке.
4. Для сравнения соответствующих дисперсий строится следующая F-статистика:

$$F = \frac{S_3/(k - m - 1)}{S_1/(k - m - 1)} = \frac{S_3}{S_1}. \quad (8.7)$$

Здесь $(k - m - 1)$ – число степеней свободы соответствующих выборочных дисперсий (m – количество объясняющих переменных в уравнении регрессии).

При сделанных предположениях относительно случайных отклонений построенная F-статистика имеет распределение Фишера с числами степеней свободы $\nu_1 = \nu_2 = k - m - 1$.

5. Если $F_{\text{набл.}} = \frac{S_3}{S_1} > F_{\text{кр.}} = F_{\alpha; \nu_1; \nu_2}$, то гипотеза об отсутствии гетероскедастичности отклоняется (здесь α – выбранный уровень значимости).

Естественным является вопрос, какими должны быть размеры подвыборок для принятия обоснованных решений. Для парной регрессии Голфелд и Квандт предлагают следующие пропорции: $n = 30$, $k = 11$; $n = 60$, $k = 22$.

Для множественной регрессии данный тест обычно проводится для той объясняющей переменной, которая в наибольшей степени связана с σ_i . При этом k должно быть больше, чем $(m + 1)$. Если нет уверенности относительно выбора переменной X_j , то данный тест может осуществляться для каждой из объясняющих переменных.

Этот же тест может быть использован при предположении об обратной пропорциональности между σ_i и значениями объясняющей переменной. При этом статистика Фишера примет вид: $F = S_1/S_3$.

8.4. Методы смягчения проблемы гетероскедастичности

Как отмечалось в разделе 8.2, гетероскедастичность приводит к неэффективности оценок, несмотря на их несмещенность. Это может привести к необоснованным выводам по качеству модели. Поэтому при установлении гетероскедастичности возникает необходимость преобразования модели с целью устранения данного недостатка. Вид преобразования зависит от того, известны или нет дисперсии σ_i^2 отклонений ε_i .

8.4.1. Метод взвешенных наименьших квадратов (ВНК)

Данный метод применяется при известных для каждого наблюдения значениях σ_i^2 . В этом случае можно устранить гетероскедастичность, разделив каждое наблюдаемое значение на соответствующее ему значение дисперсии. В этом суть метода взвешенных наименьших квадратов.

Для простоты изложения опишем ВНК на примере парной регрессии:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i. \quad (8.8)$$

Разделим обе части (9.7) на известное $\sigma_i = \sqrt{y_i^2}$:

$$\frac{y_i}{y_i} = \beta_0 \frac{1}{y_i} + \beta_1 \frac{x_i}{y_i} + \frac{\varepsilon_i}{y_i}. \quad (8.9)$$

Положив $\frac{y_i}{y_i} = y_i^*$, $\frac{x_i}{y_i} = x_i^*$, $\frac{\varepsilon_i}{y_i} = v_i$, $\frac{1}{y_i} = z_i$, получим уравнение регрессии без свободного члена, но с дополнительной объясняющей переменной Z и с “преобразованным” отклонением v :

$$y_i^* = \beta_0 z_i + \beta_1 x_i^* + v_i. \quad (8.10)$$

При этом для v_i выполняется условие гомоскедастичности. Действительно,

$$y_i^2(v_i) = M(v_i - M(v_i))^2 = M(v_i^2) - M^2(v_i).$$

Так как по предпосылке I^0 МНК $M(\varepsilon_i) = 0$, то $M(v_i) = \frac{1}{y_i^2} M(\varepsilon_i) = 0$, и

тогда $y_i^2(v_i) = M(v_i^2) =$

$$= M\left(\frac{\varepsilon_i^2}{y_i^2}\right) = \frac{1}{y_i^2} M(\varepsilon_i^2) = \frac{1}{y_i^2} M(\varepsilon_i - M(\varepsilon_i))^2 = \frac{1}{y_i^2} y_i^2 = 1 = \text{const.}$$

Следовательно, для преобразованной модели (8.10) выполняются предпосылки $1^0 - 5^0$ МНК. В этом случае оценки, полученные по МНК, будут наилучшими линейными несмещенными оценками.

Таким образом, метод взвешенных наименьших квадратов включает следующие этапы:

1. Каждую из пар наблюдений (x_i, y_i) делят на известную величину σ_i . Тем самым наблюдениям с наименьшими дисперсиями придаются наибольшие “веса”, а с максимальными дисперсиями – наименьшие “веса”. Действительно, наблюдения с меньшими дисперсиями отклонений будут более значимыми при оценке коэффициентов регрессии, чем наблюдения с большими дисперсиями. Учет этого факта увеличивает вероятность получения более точных оценок.

2. По МНК для преобразованных значений $\left(\frac{1}{y_i}, \frac{x_i}{y_i}, \frac{y_i}{y_i} \right)$ строится

уравнение регрессии без свободного члена с гарантированными качествами оценок.

8.4.2. Дисперсии отклонений не известны

Для применения ВНК необходимо знать фактические значения дисперсий y_i^2 отклонений. На практике такие значения известны крайне редко. Следовательно, чтобы применить ВНК, необходимо сделать реалистические предположения о значениях y_i^2 .

Например, может оказаться целесообразным предположить, что дисперсии y_i^2 отклонений ε_i пропорциональны значениям x_i (рис.8.5, а) или значениям x_i^2 (рис. 8.5, б).

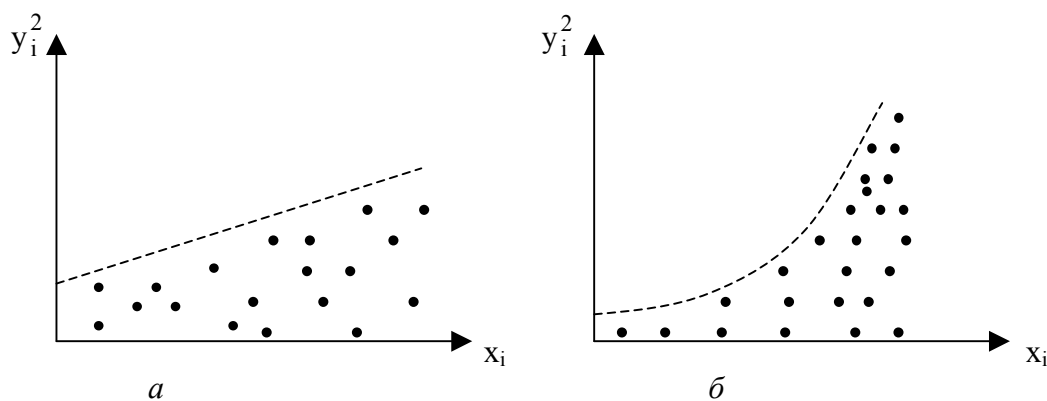


Рис. 8.5

1. Дисперсии σ_i^2 пропорциональны x_i (рис. 8.5, а).

$$y_i^2 = \sigma^2 \cdot x_i \quad (\sigma^2 - \text{коэффициент пропорциональности}).$$

Тогда уравнение (8.9) преобразуется делением его левой и правой частей на $\sqrt{x_i}$:

$$\frac{y_i}{\sqrt{x_i}} = \frac{a}{\sqrt{x_i}} + b \frac{x_i}{\sqrt{x_i}} + \frac{e_i}{\sqrt{x_i}} \quad \Rightarrow \quad \frac{y_i}{\sqrt{x_i}} = a \frac{1}{\sqrt{x_i}} + b \sqrt{x_i} + v_i. \quad (8.11)$$

Несложно показать, что для случайных отклонений $v_i = \frac{e_i}{\sqrt{x_i}}$ выполняется условие гомоскедастичности.

Следовательно, для регрессии (8.11) применим обычный МНК. Действительно, в силу выполнимости предпосылки $y_i^2 = \sigma^2(\varepsilon_i) = \sigma^2 \cdot x_i$ имеем:

$$y^2(v_i) = y^2\left(\frac{e_i}{\sqrt{x_i}}\right) = \frac{1}{x_i} y^2(e_i) = \frac{1}{x_i} y^2 \cdot x_i = y^2 = \text{const.}$$

Таким образом, оценив для (8.11) по МНК коэффициенты β_0 и β_1 , затем возвращаются к исходному уравнению регрессии (8.8).

Если в уравнении регрессии присутствует несколько объясняющих переменных, можно поступить следующим образом. Вместо конкретной объясняющей переменной X_j используется \hat{Y} исходного уравнения множественной линейной регрессии $\hat{Y} = b_0 + b_1 X_1 + \dots + b_m X_m$, т. е. фактически линейная комбинация объясняющих переменных. В этом случае получают следующую регрессию:

$$\frac{y_i}{\sqrt{\hat{Y}_i}} = b_0 \frac{1}{\sqrt{\hat{Y}_i}} + b_1 \frac{x_{i1}}{\sqrt{\hat{Y}_i}} + \dots + b_m \frac{x_{im}}{\sqrt{\hat{Y}_i}} + \frac{e_i}{\sqrt{\hat{Y}_i}}. \quad (8.12)$$

Иногда из всех объясняющих переменных выбирается наиболее подходящая, исходя из графического представления (рис. 8.4).

2. Дисперсия σ_i^2 пропорциональна x_i^2 (рис. 8.4, б).

В случае, если зависимость σ_i^2 от x_i целесообразнее выразить не линейной функцией, а квадратичной, то соответствующим преобразованием будет деление уравнения регрессии (8.8) на x_i :

$$\frac{y_i}{x_i} = b_0 \frac{1}{x_i} + b_1 + \frac{e_i}{x_i} \quad \Rightarrow \quad \frac{y_i}{x_i} = b_0 \frac{1}{x_i} + b_1 + v_i, \quad \text{где } v_i = \frac{e_i}{x_i}. \quad (8.13)$$

По аналогии с вышеизложенным несложно показать, что для отклонений v_i будет выполняться условие гомоскедастичности. После определения по МНК оценок коэффициентов β_0 и β_1 для уравнения (8.13) возвращаются к исходному уравнению (8.8).

Отметим, что для применения описанных выше преобразований существенную роль играют знания об истинных значениях дисперсий отклонений σ_i^2 , либо предположения, какими эти дисперсии могут быть. Во многих случаях дисперсии отклонений зависят не от включенных в уравнение регрессии объясняющих переменных, а от тех, которые не включены в модель, но играют существенную роль в исследуемой зависимости. В этом случае они должны быть включены в модель. В ряде случаев для устранения гетероскедастичности необходимо изменить спецификацию модели (например, линейную на лог-линейную, мультипликативную на аддитивную и т. п.).

В заключение отметим, что наличие гетероскедастичности не позволяет получить эффективные оценки, что зачастую приводит к необоснованным выводам по их качеству. Обнаружение гетероскедастичности - достаточно трудоемкая проблема и для ее решения разработано несколько методов (тестов). В случае установления наличия гетероскедастичности ее корректировка также представляет довольно серьезную проблему. Одним из возможных решений является метод взвешенных наименьших квадратов (при этом необходима определенная информация либо обоснованные предположения о величинах дисперсий отклонений). На практике имеет смысл попробовать несколько методов определения гетероскедастичности и способов ее корректировки (преобразований, стабилизирующих дисперсию).

Вопросы для самопроверки

1. В чем суть гетероскедастичности?
2. Какое из следующих утверждений верно, ложно или не определено:
 - а) вследствие гетероскедастичности оценки перестают быть эффективными и состоятельными;
 - б) оценки и дисперсии оценок остаются несмещенными;
 - в) выводы по t- и F-статистикам являются ненадежными;
 - г) при наличии гетероскедастичности стандартные ошибки оценок будут заниженными;
 - д) гетероскедастичность проявляется через низкое значение статистики Дарбина–Уотсона DW;
 - е) не существует общего теста для анализа гетероскедастичности;
 - ж) тест ранговой корреляции Спирмена основан на использовании t-статистики;
 - з) тест Парка является частным случаем теста Глейзера;
 - и) использование метода взвешенных наименьших квадратов носит ограниченный характер, т. к. для его использования необходимо знать дисперсии отклонений;

- к) если в парной регрессии дисперсия случайных отклонений пропорциональна величине объясняющей переменной (x), то для получения эффективных оценок необходимо все наблюдаемые значения поделить на x .
3. Приведите аргументы в пользу графического теста, теста Парка и теста Глейзера.
 4. Приведите схему теста Голдфелда–Квандта.
 5. В чем суть метода взвешенных наименьших квадратов (ВНК)?
 6. Объясните кратко, почему при наличии гетероскедастичности ВНК позволяет получить более эффективные оценки, чем обычный МНК.
 7. Есть основание считать, что в регрессии, построенной по квартальным данным, случайные отклонения в первых кварталах больше, нежели отклонения в других кварталах. Как это можно проверить?

Упражнения и задачи

1. Пусть зависимость заработной платы (Y) от стажа работы (X) сотрудника выражена следующим уравнением регрессии:

$$Y = \beta_0 + \beta_1 X + \gamma D + \varepsilon,$$

где D – фиктивная переменная, отражающая пол сотрудника. Как можно проверить предположение о том, что пол сотрудника не влияет на дисперсию случайных отклонений ε_i ?

2. Приведены данные в условных единицах по доходам (X) и расходам на продовольственные товары (Y) для тридцати домохозяйств:

X	26.2	33.1	42.5	47.0	48.5	49.0	49.1	50.9	52.4	53.2
Y	10.0	11.2	15.0	20.5	21.2	19.5	23.0	19.0	19.5	18.0

X	54.0	54.8	59.0	61.3	62.5	63.1	64.0	66.2	70.0	71.5
Y	24.5	21.5	35.4	25.0	17.3	21.6	15.3	32.6	34.0	23.8

X	73.2	75.4	76.0	80.6	81.2	83.3	92.0	95.5	103.2	110.4
Y	22.5	27.4	40.0	23.5	20.0	40.1	15.5	39.0	47.4	21.3

- а) Определите по МНК оценки парного уравнения регрессии $y_i = b_0 + b_1 x_i + \varepsilon_i$.
- б) Оцените качество построенного уравнения.
- в) Проведите графический анализ остатков.
- г) Примените для указанных статистических данных ВНК предположение, что $\sigma^2(\varepsilon_i) = \sigma^2 x_i^2$.
- д) Примените к полученным в п. а) результатам тест ранговой корреляции Спирмена и тест Парка.
- е) Определите, существенно ли повлияла гетероскедастичность на качество оценок в уравнении, построенном по МНК.

3. Для предприятий некоторой отрасли анализируют зависимость заработной платы (Y) сотрудников в зависимости от масштаба (от количества сотрудников) предприятия (X). Наблюдения по тридцати случайно отобраным предприятиям представлены следующей таблицей:

Y						X
75.5	75.5	77.5	78.5	80.0	81.0	100
80.5	82.0	84.5	85.0	85.5	86.5	200
85.5	88.5	90.0	91.0	95.0	96.0	300
93.0	93.5	97.5	99.0	102.5	105.0	400
102.0	105.5	107.0	110.5	115.0	118.5	500

- а) Постройте уравнение регрессии Y на X и оцените его качество.
 б) Можно ли ожидать наличие гетероскедастичности в данном случае. Ответ поясните.
 в) Проверьте наличие гетероскедастичности, используя тест Голдфелда–Квандта. Рекомендуется использовать разбиение, при котором $k = 12$.
 г) Если предположить, что гетероскедастичность имеет место, и дисперсии отклонений пропорциональны значениям X, то какое преобразование вы предложите, чтобы получить несмещенные, эффективные и состоятельные оценки.
 д) Постройте новое уравнение регрессии на основе преобразования, осуществленного в предыдущем пункте, и оцените его качество.
 е) Сравните результаты, полученные в пунктах а) и д).
4. Пусть для эмпирического уравнения парной регрессии $Y = b_0 + b_1X + e$ имеет место следующее соотношение $M(e_i^2) = \sigma^2 x_i$. Какое преобразование можно предложить, чтобы устранить проблему гетероскедастичности. Опишите поэтапно предложенную схему.
5. Пусть для регрессии $Y = b_0 + b_1X_1 + b_2X_2 + e$, оцениваемой по ежегодным данным (1971–1998), получены следующие результаты: сумма квадратов отклонений для данных 1971–1980 гг. равна $S_1 = \sum e_i^2 = 15$, для данных 1981–1998 гг. эта сумма равна $S_2 = \sum e_i^2 = 50$. С помощью теста Голдфелда–Квандта проверьте предположение о том, что дисперсия отклонений не постоянна (в частности, что дисперсия претерпела изменение где-то в 1981 г.).
6. Анализируется объем инвестиций для вымышленной страны. По данным с 1961 по 1990 г. построены два уравнения регрессии:

$$1) \hat{i}_t = 52.5 + 0.275\text{gnp}_t - 0.63c_t, \\ (t) = (12.5) \quad (10.2) \quad (6.4) \quad R^2 = 0.98.$$

$$2) \frac{\hat{i}_t}{\text{gnp}_t} = 50.7 \frac{1}{\text{gnp}_t} + 0.27 - 0.62 \frac{c_t}{\text{gnp}_t}, \\ (t) \quad (13.3) \quad (9.3) \quad (6.9) \quad R^2 = 0.87,$$

где GNP – валовой национальный продукт; C – совокупное частное потребление; I – объем инвестиций; g_{np_t} , c_t , i_t – значения соответствующих показателей в момент времени t.

- а) Что могло послужить причиной преобразования первого уравнения во второе?
- б) Если причиной преобразования являлась гетероскедастичность, то какое предположение о дисперсии отклонений являлось основанием для данного преобразования?
- в) Можно ли сравнить качества обоих уравнений на основе коэффициентов детерминации? Ответ поясните.
- г) Должно ли преобразованное уравнение проходить через начало координат?

7. Выдвигается предположение, что средняя заработная плата наемных рабочих пропорциональна их стажу. Для анализа данного утверждения обследуются по 20 рабочих восьми категорий стажа. Получены следующие статистические данные:

Стаж	[0, 5)	[5, 10)	[10, 15)	[15, 20)	[20, 25)	[25, 30)	[30, 35)	[35, 40]
З/п	10000	12500	14300	18700	25400	29000	32000	34300

- а) Постройте эмпирическое уравнение регрессии, в котором заработная плата является зависимой переменной, а стаж работы – объясняющей переменной (уравнение строится в предположение, что дисперсии отклонений постоянны).
- б) Оцените качество построенной регрессии.
- в) Есть ли основания считать, что для данной регрессионной модели весьма вероятна гетероскедастичность? Если да, то почему?
- г) Предполагая, что дисперсия отклонений пропорциональна трудовому стажу, постройте на основании тех же данных уравнение по методу взвешенных наименьших квадратов (ВНК).
- д) Предполагая, что дисперсия отклонений пропорциональна квадрату величины трудового стажа, постройте по ВНК соответствующее уравнение регрессии.
- е) Какое из трех предположений относительно дисперсии отклонений наиболее реалистично с вашей точки зрения?

8. Исследуется зависимость между доходом (X) домохозяйства и его расходом (Y) на продукты питания. Выборочные данные по 40 домохозяйствам представлены ниже.

X	25.5	26.5	27.2	29.6	35.7	38.6	39.0	39.3	40.0	41.9	42.5	44.2	44.8	45.5
Y	14.5	11.3	14.7	10.2	13.5	9.9	12.4	8.6	10.3	13.9	14.9	11.6	21.5	10.8
X	45.5	48.3	49.5	52.3	55.7	59.0	61.0	61.7	62.5	64.7	69.7	71.2	73.8	74.7
Y	13.8	16.0	18.2	19.1	16.3	17.5	10.9	16.1	10.5	10.6	29.0	8.2	14.3	21.8

X	75.8	76.9	79.2	81.5	82.4	82.8	83.0	85.9	86.4	86.9	88.3	89.0
Y	26.1	20.0	19.8	21.2	29.0	17.3	23.5	22.0	18.3	13.7	14.5	27.3

- Постройте эмпирическое уравнение регрессии Y на X.
- Вычислите отклонения e_i .
- Проведите анализ модели на гетероскедастичность по тесту ранговой корреляции Спирмена.
- Проведите графический анализ отклонений и выдвиньте предположение о зависимости дисперсии отклонений от значений X.
- На основании предыдущего пункта постройте новое уравнение регрессии, используя для этого ВНК.

9. Проводится анализ зависимости средней заработной платы от средней производительности на предприятиях различного масштаба. Проведенное обследование нашло отражение в следующей таблице.

Количество сотрудников предприятия, n	Средняя производительность, X (\$)	Средняя з/п, Y (\$)	Стандартное отклонение з/п, σ_i (\$)
1 – 4	9320	3320	740
4 – 9	8630	3640	850
10 – 19	8050	3900	730
20 – 49	8320	4120	820
50 – 99	8600	4090	950
100 – 199	9120	4200	1100
200 – 499	9540	4380	1250
500 – 999	9730	4500	1290
1000 – 1999	10120	4610	1350
2000 – 4999	10740	4800	1100
> 5000	11200	5000	1520

- Постройте уравнение регрессии $y_i = b_0 + b_1x_i + e_i$, используя обычный МНК.

- Постройте уравнение регрессии $\frac{y_i}{Y_i} = b_0 \frac{1}{Y_i} + b_1 \frac{x_i}{Y_i} + \frac{e_i}{Y_i}$.

- Сравните полученные результаты. Какое из уравнений вы предпочтете и почему?