Лекция 4
Обобщенная модель множественной регрессии

1. Мультиколлинеарность

2. Гетероскедастичность

1. Мультиколлинеарность

Под *мультиколлинеарностью* понимают высокую взаимную коррелированность объясняющих переменных.

Мультиколлинеарность может проявляться в *функциональной* и *стохастической* формах.

В первом случае, по крайней мере, одна пара из объясняющих переменных связана линейной функциональной зависимостью и тогда говорят о *строгой мультиколлинеарности* этих факторов. В этом случае в матрице X в силу линейной зависимости двух её столбцов нарушается предпосылка 6° МНК – ранг матрицы X будет меньше, чем $p+1$.

В этом случае матрица $X'X$ будет ***вырожденной*** и обратной матрицы $(X'X)^{-1}$ просто не существует. Оценку параметров модели невозможно найти из нормального векторного уравнения

$$X'Xb = X'Y.$$

На практике строгая мультиколлинеарность встречается достаточно редко, т. к. её несложно избежать на стадии предварительного отбора факторов модели.

Чаще связь между объясняющими переменными выражается в стохастической форме, когда они тесно коррелируют друг с другом.

В этом случае говорят о *нестрогой мультиколлинеарности*.

Матрица $X'X$ хотя и неособенная, но её определитель $|X'X|$ близок к нулю. Компоненты вектора оценок b обратно пропорциональны величине определителя и в силу этого имеют значительные средние квадратические отклонения σ_{b_j}, и, следовательно, большие стандартные ошибки m_{b_j}. Отсюда они нестабильны как по величине, так и по знаку.

В итоге отметим основные *негативные* последствия мультиколлинеарности:

- большие дисперсии оценок параметров $D(b_j)$ приводят к существенным отклонениям оценок от оцениваемого параметра, расширяет интервальные оценки;

- уменьшаются $t-$ статистики параметров , что может привести к неоправданному выводу о статистической незначимости параметров b_j и о несущественном влиянии соответствующего фактора на результат y;

- МНК- оценки b_j коэффициентов модели и их стандартные ошибки m_{b_j} становятся очень чувствительными к малейшему изменению исходных данных;
- становится невозможным определить изолированное влияние факторов на результат y .

Точных количественных критериев для установления или отсутствия мультиколлинеарности не существует. Но существуют некоторые *эвристические* подходы к её выявлению.

Один из них заключается в анализе матрицы межфакторной корреляции

$$R_x = \begin{pmatrix} 1 & r_{x_1x_2} & r_{x_1x_3} & \cdots & r_{x_1x_p} \\ r_{x_2x_1} & 1 & r_{x_2x_3} & \cdots & r_{x_2x_p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{x_px_1} & r_{x_px_2} & r_{x_px_3} & \cdots & 1 \end{pmatrix}$$

Считается, что если в ней содержатся коэффициенты корреляции, у которых $\left|r_{x_i x_j}\right| \geq 0{,}75$, то это свидетельствует о присутствии *нестрогой* мультиколлинеарности.

Другой подход в оценке мультиколлинеарности состоит в исследовании определителя матрицы $X'X$. Если $|X'X| = 0$, то существует *строгая* мультиколлинеарность, а если он близок к нулю ($|X'X| < 0{,}1$), то это свидетельствует о наличии *нестрогой* мультиколлинеарности.

Для оценки значимости мультиколлинеарности факторов можно использовать определитель матрицы межфакторной корреляции $|R_x|$. Если бы факторы не коррелировали между собой, то все внедиагональные элементы матрицы R_x равнялись бы нулю. Если же все $r_{x_i x_j} = 1$, то определитель такой матрицы равнялся бы нулю.

Отсюда выдвигается гипотеза $H_0 : |R_x| = 1$ (отсутствие мультиколлинеарности).
Доказано, что статистика

$$\chi^2 = n - 1 - 1/6(2p + 5)\lg|R_x|$$

имеет приближенное распределение «хи-квадрат» с $k = 1/2n(n-1)$ степенями свободы.

Если $\chi^2_{набл} > \chi^2_{кр}(\alpha, k)$, то гипотеза H_0 отклоняется и мультиколлинеарность факторов считается доказанной.

Если мультиколлинеарность установлена, то каким образом её можно устранить?

Единого подхода к её устранению не существует, но используются ряд методов, которые применимы в конкретных ситуациях.

Самый простой из них заключается в том, что из двух объясняющих переменных, имеющих высокий коэффициент корреляции ($\left|r_{x_i x_j}\right| \geq 0{,}75$), одну из переменных исключают из уравнения.

Но здесь нужна осторожность, чтобы не исключить переменную, которая ***необходима*** в уравнении по своей экономической сущности, но зачастую коррелирует с другими факторами.

Другой метод заключается в увеличении объёма выборки, если это возможно: большее количество данных позволяет получить МНК-оценки с меньшей дисперсией.

Например, при использовании ежегодных данных можно перейти к поквартальным данным и объем выборки увеличится в 4 раза.

В следующем методе переходят от *несмещенных* МНК-оценок параметров к таким *смещенным* оценкам, которые обладают меньшим рассеиванием относительно математического ожидания (рис. 1).

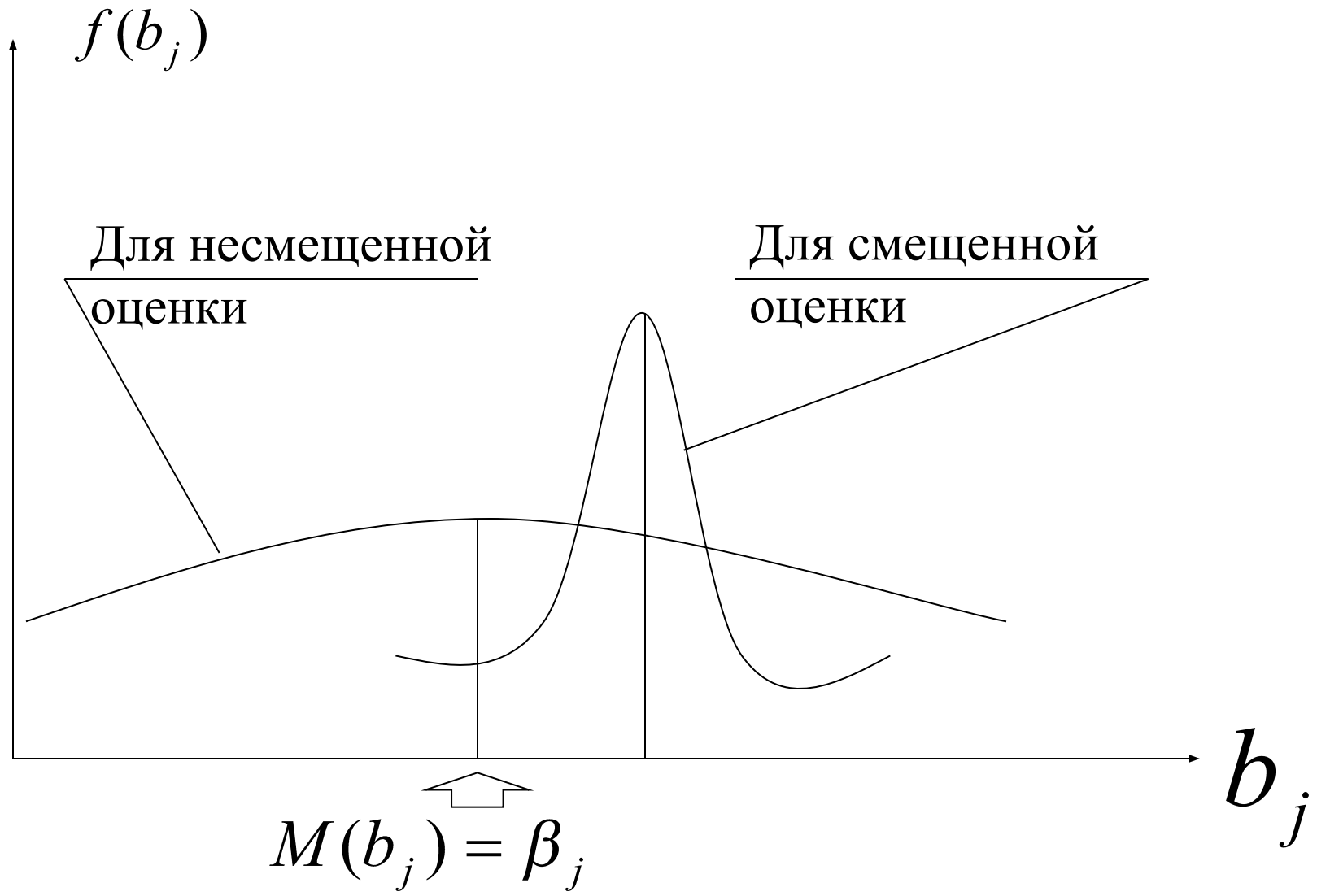
$f(b_j)$
Для несмещенной оценки
Для смещенной оценки
b_j
$M(b_j) = \beta_j$

Рис. 1

Например, при использовании «ридж-регрессии» (гребневой регрессии) рассматривают смещенные оценки

$$b_\tau = \left(X'X + \tau E_{p+1}\right)^{-1} X'Y,$$

где τ – некоторое малое положительное число называемое *гребнем*, E_{p+1} – единичная матрица порядка $p+1$. Диагональные элементы матрицы $X'X$ при этом увеличиваются на величину τ , а остальные элементы остаются неизменными.

Определитель матрицы $\left|X'X+\tau\cdot E_{p+1}\right|$ увеличивается по сравнению с $\left|X'X\right|$ и эффект мультиколлинеарности уменьшается.

При плохой обусловленности матрицы для оценки параметров иногда применяют *метод главных компонент*. Основная идея метода состоит в замене исходных объясняющих переменных x_j, $j=\overline{1,p}$ на новые переменные z_i, $i=\overline{1,k}$. Новые переменные (главные компоненты) должны обладать следующими свойствами:

- полная совокупность главных компонент должна содержать в себе всю изменчивость исходных переменных

$$x_j,\ j = \overline{1, p};$$

- главные компоненты должны быть ортогональны между собой, т. е. быть линейно-независимыми.

2. Гетероскедастичность

Предпосылка 3° МНК о постоянстве дисперсий $D(\varepsilon_i)$ случайных составляющих ε_i для всех наблюдений на практике не всегда выполняется и имеет место *гетероскедастичность* модели.

Негативные последствия гетероскедастичности следующие:

- оценки коэффициентов модели, оставаясь несмещенными и состоятельными, уже не будут *эффективными*, и при небольших объёмах выборок появляется риск получения оценок b_j, существенно отличающихся от оцениваемого коэффициента β_j;

- стандартные ошибки параметров , как правило, будут заниженными, а t -статистики – завышенными, что приводит к признанию статистической значимости параметров, которые на самом деле таковыми не являются;
- дисперсии оценок $D(b_j)$ будут рассчитываться со смещением, что существенно влияет на интервальные оценки коэффициентов модели.

Для обнаружения гетероскедастичности наиболее простым является *визуальный* метод.

Наличие гетероскедастичности для ***парной*** регрессии можно наглядно видеть из поля корреляции, когда дисперсия случайных составляющих растет (или уменьшается) по мере увеличения x (рис. 2).

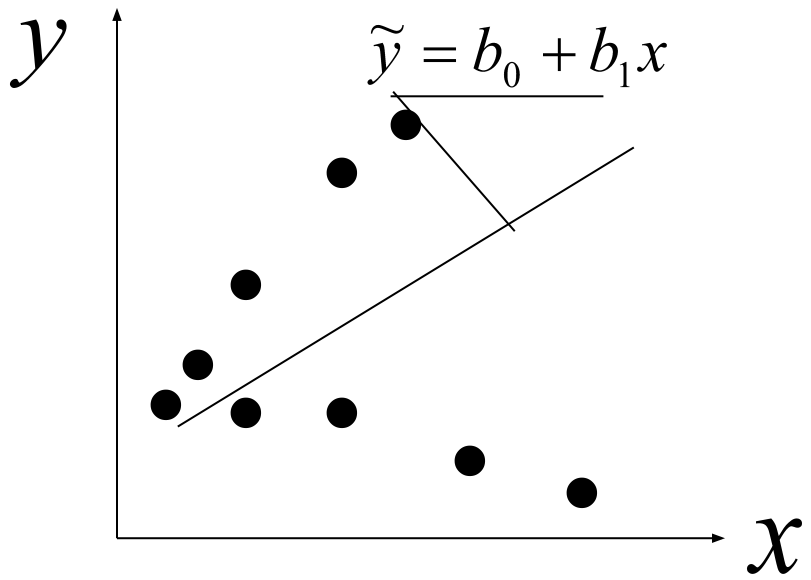

Рис. 2

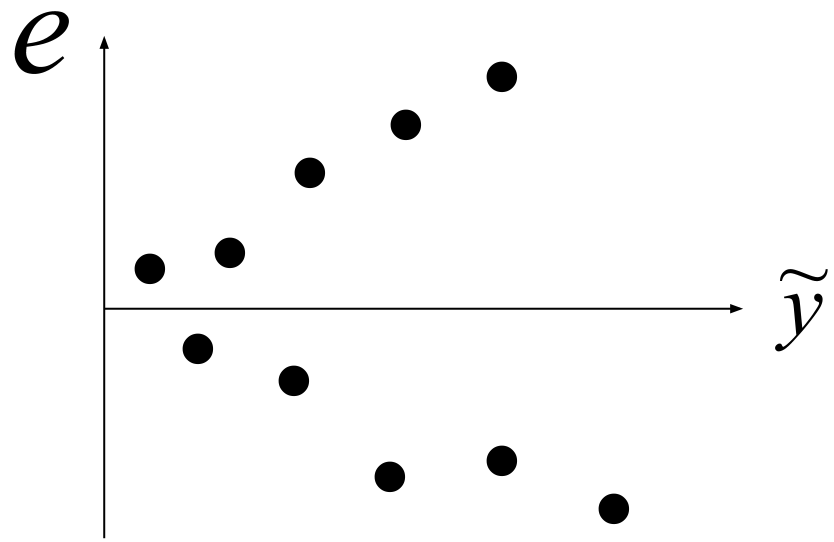

Рис. 3

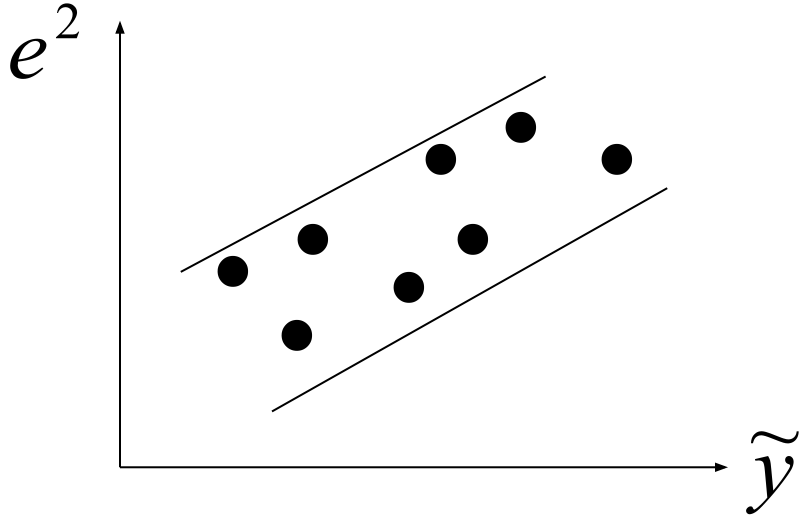

Рис. 4

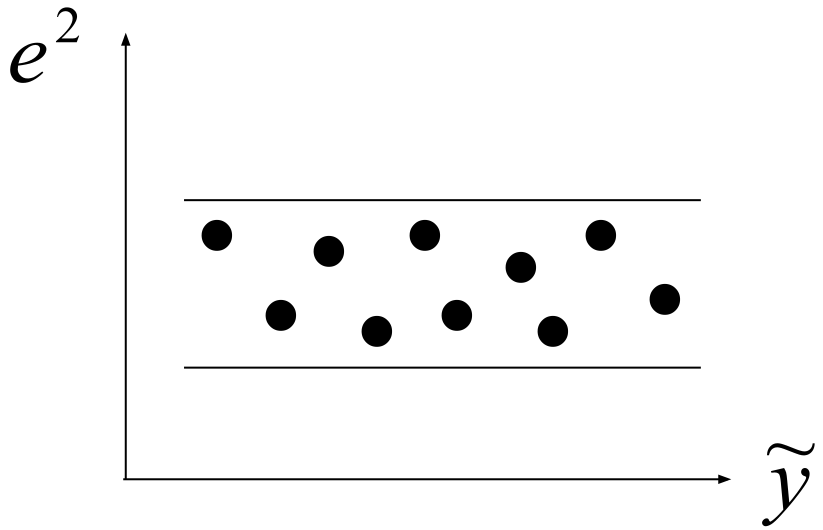

Рис. 5

В некоторых случаях гетероскедастичность визуально не столь очевидна. Тогда применяют тесты на гетероскедастичность, причем все они используют нулевую гипотезу об *отсутствии* гетероскедастичности.

Тест *ранговой корреляции Спирмена* использует наиболее общее предположение о зависимости дисперсий ошибок от значений объясняющей переменной x :

$$D(\varepsilon_i) = \sigma_i^2 = f_i(x_i), i = \overline{1,n}.$$

Никаких дополнительных предположений относительно вида функций f_i и законе распределения возмущений ε_i здесь не делается.

Идея теста заключается в том, что $|e_i|$ является некоторой оценкой σ_i, и поэтому в случае гетероскедастичности значения $|e_i|$ будут коррелировать x_i.

Рассмотрим применение теста на примере парной регрессии $\tilde{y} = b_0 + b_1 x$. В тесте используют коэффициент ранговой корреляции r_{xe} , для нахождения которого следует отдельно ранжировать наблюдения по возрастанию переменной x_i , когда каждое значение x_i получит свой ранг от 1 до n , а таким же образом ранжировать остатки $|e_i|$.

В итоге коэффициент r_{xe} вычисляется по формуле:

$$r_{xe} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \qquad (1)$$

где $d_i -$ разность между рангами x_i и $|e_i|$.

Доказано, что при справедливости гипотезы $H_0 : r_{xe} = 0$ статистика

$$T = \frac{r_{xe} \cdot \sqrt{n-2}}{\sqrt{1-r_{xe}^2}} \qquad (2)$$

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

Поэтому, если $|t_{набл}|$ превышает $t_{кр}(\frac{\alpha}{2}, k)$, то гипотезу H_0 отклоняют и признают **наличие** гетероскедастичности.

Для множественной регрессии проверка гипотезы H_0 с помощью статис-тики (2) может выполняться по каждому фактору **отдельно**.

Тест *Голдфельда-Квандта* применяется в том случае, когда случайные величины ε_i имеют нормальное распределение и

$$\operatorname{cov}(\varepsilon_i, \varepsilon_j) = 0, i \neq j \quad .$$

В нём предполагается, что дисперсии σ_i^2 возмущения ε_i пропорциональны квадрату переменной x_i, т. е.

$$\sigma_i^2 = \sigma^2 x_i^2 \quad .$$

На примере парной регрессии

$$\tilde{y} = b_0 + b_1 x$$

тест состоит из следующих этапов.

1. Все наблюдений упорядочиваются в порядке возрастания переменной x .
2. Вся упорядоченная выборка разбивается на три подвыборки объёмов соответст-венно (обычно $l,\quad n-2l,\ ).\ l$ $\qquad\qquad l \approx n/3$
3. Оцениваются отдельно две регрессии

$$\tilde{y} = b_{10} + b_{11}x, \quad \tilde{y} = b_{20} + b_{21}x$$

для первой подвыборки (первые l наблю-дений) и третьей подвыборки (последние l наблюдений).

Рассчитываются остаточные суммы для обеих регрессий

$$s_1 = \sum_{i=1}^{l} e_i^2, \quad s_2 = \sum_{i=n-l+1}^{n} e_i^2.$$

4. Выдвигается гипотеза

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2,$$

для проверки которой используется статистика

$$F = \begin{cases} \dfrac{s_2}{s_1}, \text{ если } s_2 > s_1, \\ \dfrac{s_1}{s_2}, \quad \text{если } s_1 > s_2, \end{cases}$$

которая при справедливости гипотезы H_0 имеет распределение Фишера с степенями $k_1 = k_2 = l - p - 1$ свободы. Если $F_{набл} > F_{кр}(\alpha, k_1, k_2)$, то гипотеза об отсутствии гетероскедастич-ности отклоняется на уровне значимости α.

Если в модели более одного фактора, то выборка упорядочивается по тому фактору, который, как предполагается, теснее связан с σ_i^2.

При установлении гетероскедастичности возникает необходимость преобразования модели с целью устранения данного недостатка.

Если дисперсии σ_i^2 *известны*, то гетероскедастичность легко *устраняется*. Рассмотрим это на примере парной регрессии

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = \overline{1,n} \qquad (3)$$

Разделим обе части уравнения (3) на известное значение

$$\sigma_i = \sqrt{\sigma_i^2}$$

и сделаем замену переменных:

$$\frac{y_i}{\sigma_i} = \beta_0 \frac{1}{\sigma_i} + \beta_1 \frac{x_i}{\sigma_i} + \frac{\varepsilon_i}{\sigma_i}$$

$$y_i^* = \frac{y_i}{\sigma_i}, \quad x_i^* = \frac{x_i}{\sigma_i}, \quad z_i = \frac{1}{\sigma_i}, \quad \nu_i = \frac{\varepsilon_i}{\sigma_i}$$

Тогда получим модельное уравнение регрессии с двумя факторами x_i^*, z_i, но без свободного члена

$$y_i^* = \beta_0 z_i + \beta_1 x_i^* + \nu_i, \quad i = \overline{1, n}. \quad (4)$$

Очевидно, что для любого наблюдения

$$D(v_i) = D(\frac{\varepsilon_i}{\sigma_i}) = \frac{1}{\sigma_i^2} D(\varepsilon_i) = \frac{\sigma_i^2}{\sigma_i^2} = 1 = const,$$

т.е. модель (4) является *гомоскедастичной*, классической.

Полученные МНК - оценки b_0, b_1 коэффициентов модели (4) будут наилучшими несмещенными оценками и их можно использовать для первоначальной модели (3).

Уравнение (4) представляет собой взвешенную регрессию с весами $1/\sigma_i$.

Наблюдения с наименьшими дисперсиями получат наибольшие "веса" и наоборот. Поэтому данную версию МНК называют *взвешенным методом наименьших квадратов* (ВМНК). В свою очередь он является частным случаем *обобщенного* метода наименьших квадратов (ОМНК), когда оценки определяются по формуле:

$$\hat{b} = \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}Y.$$

Здесь Ω – ковариационная положительно определенная матрица ошибок, т.е. $\Sigma_\varepsilon = \Omega$ и её диагональные элементы различны, а внедиагональные элементы в общем случае не равны нулю (в классической модели Σ_ε представляет скалярную матрицу с одинаковыми диагональными элементами σ^2).

На практике значения σ_i^2 неизвестны. Поэтому, чтобы применить ВМНК, необходимо сделать реалистические предположения о значениях σ_i^2. В этих случаях говорят не об устранении, а о *смягчении* гетероскедастичности.

Если предположить, что дисперсии σ_i^2 пропорциональны значениям x_i

$$\sigma_i^2 = \sigma^2 x_i, \quad i = \overline{1, n},$$

тогда уравнение (3) преобразуется в гомоскедастичную модель делением обеих его частей на $\sqrt{x_i}$:

$$y_i^* = \beta_0 z_i + \beta_1 x_i^* + \nu_i,$$

где

$$y_i^* = \frac{y_i}{\sqrt{x_i}}, \quad x_i^* = \frac{x_i}{\sqrt{x_i}}, \quad z_i = \frac{1}{\sqrt{x_i}}, \quad \nu_i = \frac{\varepsilon_i}{\sqrt{x_i}}.$$

Если же предположить, что дисперсии σ_i^2 пропорциональны значениям квадратов x_i

$$\sigma_i^2 = \sigma^2 x_i^2, \quad i = \overline{1, n},$$

то делением обеих его частей на величину x_i можно получить гомоскедастичную модель

$$y_i^* = \beta_0 z_i + \beta_1 + \nu_i.$$

Отметим, что параметры в последней модели по сравнению с уравнением (3) поменялись ролями: β_0 – коэффициент регрессии, β_1 – свободный член.