

# II. Метод максимального правдоподобия в эконометрии

## Оглавление

БАЗОВЫЕ ПОНЯТИЯ .....	3
ХАРАКТЕРИСТИКА ММП.....	9
Связь ММП с МНК. Квази-МП методы.....	11
Связь ГЕССИАНА И МАТРИЦЫ ВКЛАДОВ В ГРАДИЕНТ С ИНФОРМАЦИОННОЙ МАТРИЦЕЙ .....	12
➤ Гессиан и информационная матрица .....	12
➤ Матрица вкладов в градиент и информационная матрица .....	13
➤ Вычисление информационной матрицы .....	14
РАСПРЕДЕЛЕНИЕ ГРАДИЕНТА И ОЦЕНОК МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ ...	16
➤ Асимптотическое распределение градиента и оценок максимального правдоподобия.....	16
➤ Выборочная оценка распределения градиента и оценок максимального правдоподобия.....	17
ЧИСЛЕННЫЕ МЕТОДЫ НАХОЖДЕНИЯ ОЦЕНОК МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ.....	20
ММП И ПРОВЕРКА ГИПОТЕЗ .....	22
➤ Асимптотическое распределение и асимптотическая эквивалентность трех классических статистик.....	22
➤ Соотношения между статистиками.....	27
МОДЕЛИ С ДИСКРЕТНОЙ ЗАВИСИМОЙ ПЕРЕМЕННОЙ .....	30
➤ Модели с бинарной зависимой переменной (логит и пробит).....	30
➤ Пуассоновская регрессия .....	33
ОБОБЩЕННЫЙ МЕТОД НАИМЕНЬШИХ КВАДРАТОВ .....	36
РЕГРЕССИИ С НЕОДИНАКОВОЙ ДИСПЕРСИЕЙ И ТЕСТИРОВАНИЕ ГЕТЕРОСКЕДАСТИЧНОСТИ .....	39
➤ Взвешенный метод наименьших квадратов.....	39
➤ Проверка гипотезы о наличии гетероскедастичности известного вида.....	40
➤ Регрессия с мультипликативной гетероскедастичностью .....	42
НЕЛИНЕЙНАЯ РЕГРЕССИЯ. МЕТОД ГАУССА-НЬЮТОНА .....	45
ОЦЕНИВАНИЕ РЕГРЕССИИ С AR-ОШИБКОЙ .....	46
➤ Нелинейная регрессия с пропущенным первым наблюдением.....	46
➤ Оценивание регрессии с AR(1)-ошибкой полным методом максимального правдоподобия.....	47
РЕГРЕССИЯ С MA-ОШИБКОЙ .....	50

➤ Оценивание регрессии с MA(1)-процессом в ошибке полным методом максимального правдоподобия .....	50
➤ Оценивание регрессии с MA-ошибкой нелинейным МНК .....	52
<b>РЕГРЕССИЯ С ARCH-ПРОЦЕССОМ В ОШИБКЕ .....</b>	<b>53</b>
<b>ЯКОБИАН ПРЕОБРАЗОВАНИЯ ПЛОТНОСТИ РАСПРЕДЕЛЕНИЯ В ФУНКЦИИ ПРАВДОПОДОБИЯ.....</b>	<b>59</b>
➤ Функция правдоподобия модели типа $\varepsilon = f(Y, \theta_1)$ .....	59
➤ Преобразование зависимой переменной. Модель Бокса-Кокса.....	60
<b>ТЕСТ НА НОРМАЛЬНОСТЬ .....</b>	<b>63</b>
<b>РЕГРЕССИЯ С ОШИБКАМИ ВО ВСЕХ ПЕРЕМЕННЫХ .....</b>	<b>67</b>
<b>ВНЕШНЕ НЕ СВЯЗАННЫЕ РЕГРЕССИОННЫЕ УРАВНЕНИЯ .....</b>	<b>70</b>
<b>СИСТЕМЫ ОДНОВРЕМЕННЫХ УРАВНЕНИЙ .....</b>	<b>75</b>
➤ FIML .....	76
➤ LIML .....	79
<b>ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА .....</b>	<b>81</b>
<b>ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ .....</b>	<b>83</b>

## Базовые понятия

Пусть  $Y$  — реализация  $N$ -мерной случайной величины, распределенной как:

а)  $P_{\theta}(x)$  (вероятность) — в случае дискретного распределения.

б)  $p_{\theta}(x)$  (плотность) — в случае непрерывного распределения.

Здесь  $P_{\theta}(x)$  ( $p_{\theta}(x)$ ) характеризует семейство распределений задаваемое параметром  $\theta \in \Theta$ ,  $\Theta \subset \mathbb{R}^m$  — пространство параметров. В эконометрии принято говорить об этом семействе распределений как о **порождающем данные процессе** (ПДП). Будем считать, что рассматриваемый вектор наблюдений (выборка) порожден распределением из этого семейства с параметром  $\theta_0 \in \Theta$ , которое будем называть истинным распределением, а  $\theta_0$  — истинным параметром.

Функция  $\mathcal{L}(Y, \theta) = P_{\theta}(Y)$  (соответственно  $\mathcal{L}(Y, \theta) = p_{\theta}(Y)$ ) называется **функцией правдоподобия**.

**Оценкой максимального правдоподобия** ( $\hat{\theta}$ ), сокращенно оценкой МП, называется решение задачи

$$\mathcal{L}(Y, \theta) \rightarrow \max_{\theta \in \Theta}.$$

Будем считать в дальнейшем, что решение задачи единственно.<sup>1</sup>

Такой метод оценивания называют **методом максимального правдоподобия**.

Обычно удобнее пользоваться **логарифмической функцией максимального правдоподобия**

$$\ell(Y, \theta) = \ln(\mathcal{L}(Y, \theta)).$$

Логарифм — (бесконечно) дифференцируемая возрастающая функция: поэтому можно находить оценки МП решая задачу  $\ell(Y, \theta) \rightarrow \max_{\theta \in \Theta}$ .

В частном случае вектор наблюдений состоит из независимых случайных величин. При этом

$$\mathcal{L}(Y, \theta) = \prod_i \mathcal{L}_i(Y_i, \theta), \quad \ell(Y, \theta) = \sum_i \ell_i(Y_i, \theta).$$

Вообще говоря вектор наблюдений  $Y$  состоит из зависимых между собой и/или неодинаково распределенных случайных величин, поэтому не является

---

<sup>1</sup> Пример неединственности представляет оценивание процесса скользящего среднего в ошибке.

выборкой в обычном смысле слова. В общем случае это равенство тоже будет верным если обозначить

$$\mathcal{L}_i(Y_i, \theta) = p_{\theta}(Y_i | Y_{i-1}, \dots, Y_1) \quad \text{и} \quad \ell_i(Y_i, \theta) = \ln(\mathcal{L}_i(Y_i, \theta)).$$

Тем самым задается разбиение функции правдоподобия на **вклады отдельных наблюдений**.

Поскольку  $Y$  — случайная величина, то функция правдоподобия — случайная величина при данном значении параметров. Оценка максимального правдоподобия является функцией вектора наблюдений:  $\hat{\theta} = \hat{\theta}(Y)$ , поэтому это тоже случайная величина. Соответственно, точно так же случайными величинами является значение функции правдоподобия в максимуме  $\hat{\mathcal{L}}(Y) = \mathcal{L}(Y, \hat{\theta})$  и многие другие рассматриваемые далее величины (градиент, гессиан и т. п.).

Пусть функция правдоподобия дифференцируема по  $\theta$  и достигает максимума во внутренней точке ( $\hat{\theta} \in \text{int}(\Theta)$ ), тогда оценка МП  $\hat{\theta}$  должна удовлетворять следующему условию первого порядка:

$$\frac{\partial \mathcal{L}}{\partial \theta}(Y, \hat{\theta}) = 0 \quad \text{или} \quad \frac{\partial \ell}{\partial \theta}(Y, \hat{\theta}) = 0.$$

Таким образом, **градиент** логарифмической функции правдоподобия  $g(\theta)$  при  $\theta = \hat{\theta}$  должен быть равен нулю.

Для того, чтобы оценки, удовлетворяющие этим **уравнениям правдоподобия** действительно давали максимум правдоподобия, необходимо и достаточно, чтобы были выполнены условия второго порядка (предполагаем, что функция правдоподобия дважды дифференцируема). А именно, **матрица Гессе (гессиан)** логарифмической функции правдоподобия должна быть всюду отрицательно определена. Далее мы встретим случаи, когда это свойство действительно выполнено (логит и пробит), и когда может быть несколько локальных максимумов (“полная” функция правдоподобия для регрессии с AR(1)-ошибкой). Матрица Гессе  $\mathcal{H}$  по определению есть матрица вторых производных:

$$\mathcal{H}_{jl}(Y, \theta) = \frac{\partial^2 \ell}{\partial \theta_j \partial \theta_l}(Y, \theta) \quad j, l = 1, \dots, m.$$

С помощью матричного дифференцирования можно записать гессиан в виде

$$\mathcal{H} = \frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}.$$

В некоторых моделях функция правдоподобия неограниченна сверху и не существует оценок максимального правдоподобия в смысле приведенного выше определения. Согласно альтернативному определению оценками максимального правдоподобия называют корни уравнения правдоподобия, являющиеся локальными максимумами функции правдоподобия, корнями уравнения правдоподобия. Существуют модели, для которых такие оценки состоятельны.

**Информационной матрицей** для вектора наблюдений размерностью  $N$  будем называть матрицу

$$\mathcal{I}^N(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(\mathbf{g}(Y, \boldsymbol{\theta}) \mathbf{g}^\top(Y, \boldsymbol{\theta})).$$

Заметим, что по этому определению информационная матрица — функция некоторого вектора параметров  $\boldsymbol{\theta} \in \Theta$ . В дальнейшем для сокращения записи, если это не вносит путаницы, индекс количества наблюдений  $N$  будем опускать:  $\mathcal{I}(\boldsymbol{\theta})$ . Индекс  $\boldsymbol{\theta}$  у символа математического ожидания  $E$  означает, что ожидание вычисляется в предположении, что  $\boldsymbol{\theta}$  — точка истинных параметров. Заметьте, что оператор  $E$  без нижнего индекса везде означает ожидание для распределения с параметрами  $\boldsymbol{\theta}_0$ !

В дальнейшем будет использоваться следующее очевидное свойство функции правдоподобия. Пусть  $\varphi(Y)$  есть некоторая функция вектора наблюдений  $Y$ . Тогда ее математическое ожидание равно

$$E(\varphi(Y)) = \int_{\mathcal{Y}} \varphi(Y) \mathcal{L}(\boldsymbol{\theta}_0, Y) dY,$$

где  $\mathcal{Y}$  обозначает пространство элементарных событий (пространство переменной  $Y$ ).

Таким образом, можно переписать определение информационной матрицы в виде

$$\mathcal{I}(\boldsymbol{\theta}) = \int_{\mathcal{Y}} \mathbf{g}(Y, \boldsymbol{\theta}) \mathbf{g}^\top(Y, \boldsymbol{\theta}) \mathcal{L}(\boldsymbol{\theta}, Y) dY.$$

**Асимптотическая информационная матрица** есть предел

$$\mathcal{I}^\infty(\boldsymbol{\theta}) = \lim_{N \rightarrow \infty} \frac{1}{N} \mathcal{I}^N(\boldsymbol{\theta}).$$

Множитель  $1/N$  добавлен в определения для того, чтобы существовал конечный предел (информационная матрица является величиной порядка  $O(N)$ ).

Если мы рассматриваем выборку, то применяя определение информационной матрицы к отдельным наблюдениям ( $\mathcal{I}_i$ ), имеем

$$\mathcal{I}^N = N\mathcal{I}_i.$$

Таким образом, если наблюдения независимы и одинаково распределены, то информация растет пропорционально количеству наблюдений.

**Пример.** Линейная регрессия с нормально распределенными ошибками.

Пусть ошибки  $\varepsilon_i \sim \text{NID}(0, \sigma^2)$ . Эта аббревиатура означает, что случайные величины  $\varepsilon_i$  независимы и имеют нормальное распределение с параметрами  $(0, \sigma^2)$  (normally and independently distributed). Ковариационная матрица вектора ошибок — это единичная матрица с точностью до множителя:  $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \mathbf{I}_N$ .

Зависимая переменная связана с ошибками следующим образом:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

где  $\mathbf{X}$  — матрица регрессоров ( $N \times m$ ),  $\boldsymbol{\beta}$  — вектор-столбец неизвестных коэффициентов длины  $m$ . Таким образом,  $Y_i$  имеет нормальное распределение с параметрами  $(\mathbf{X}_i\boldsymbol{\beta}, \sigma^2)$ , где  $\mathbf{X}_i$  —  $i$ -я строка матрицы  $\mathbf{X}$ :

$$Y_i \sim \text{N}(\mathbf{X}_i\boldsymbol{\beta}, \sigma^2).$$

Плотность распределения  $\text{N}(\alpha, \sigma^2)$  равна

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \alpha)^2}{2\sigma^2}\right).$$

Функция правдоподобия для этого набора наблюдений имеет вид

$$\mathcal{L} = (2\pi\sigma^2)^{-N/2} \prod_{i=1}^N \exp\left(-\frac{(Y_i - \mathbf{X}_i\boldsymbol{\beta})^2}{2\sigma^2}\right).$$

Логарифмическая функция правдоподобия:

$$\begin{aligned} \ell &= -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^N (Y_i - \mathbf{X}_i\boldsymbol{\beta})^2 = \\ &= -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \mathbf{e}^\top \mathbf{e}. \end{aligned}$$

Здесь мы обозначили вектор остатков  $\mathbf{e} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$ .

В данном случае вектор неизвестных параметров состоит из двух компонент:

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\beta} \\ \sigma^2 \end{pmatrix}.$$

Градиент логарифмической функции правдоподобия тоже состоит из двух частей:

$$\mathbf{g}_{\boldsymbol{\beta}} = \frac{\partial \ell}{\partial \boldsymbol{\beta}^\top} = \frac{1}{\sigma^2} \mathbf{X}^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{e}.$$

$$g_{\sigma^2} = \frac{\partial \ell}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{\mathbf{e}^\top \mathbf{e}}{2\sigma^4} = \frac{1}{2\sigma^4} (\text{RSS}(\boldsymbol{\beta}) - N\sigma^2),$$

где  $\text{RSS}(\boldsymbol{\beta}) = \mathbf{e}^\top \mathbf{e}$  — сумма квадратов остатков.

Оценка максимального правдоподобия  $\hat{\boldsymbol{\theta}}$  должна удовлетворять равенству  $\mathbf{g}(\hat{\boldsymbol{\theta}}) = 0$ , откуда получим

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \quad \text{и} \quad \hat{\sigma}^2 = \frac{\text{RSS}(\hat{\boldsymbol{\beta}})}{N} = \frac{1}{N} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

ММП дает ту же оценку вектора коэффициентов регрессии  $\boldsymbol{\beta}$ , что и МНК. Как известно, оценка дисперсии  $\hat{\sigma}^2$  является смещенной:

$$E(\hat{\sigma}^2) = \frac{N-m}{N} \sigma^2.$$

Покажем, каким образом связаны ММП и МНК.

Выразим, используя равенство  $g_{\sigma^2} = 0$ , дисперсию через  $\boldsymbol{\beta}$ :

$$\sigma^2(\boldsymbol{\beta}) = \frac{\text{RSS}(\boldsymbol{\beta})}{N}.$$

Если подставить ее в функцию правдоподобия, то получится **концентрированная функция правдоподобия**:

$$\ell^c = -\frac{N}{2} \ln(2\pi\sigma^2(\boldsymbol{\beta})) - \frac{1}{2\sigma^2(\boldsymbol{\beta})} \text{RSS}(\boldsymbol{\beta}) = -\frac{N}{2} \ln\left(2\pi \frac{\text{RSS}(\boldsymbol{\beta})}{N}\right) - \frac{N}{2}.$$

Максимизация ее эквивалентна минимизации суммы квадратов остатков  $\text{RSS}(\boldsymbol{\beta})$  по  $\boldsymbol{\beta}$ .

Гессиан логарифмической функции правдоподобия состоит из следующих компонент:

$$\mathcal{H}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = -\frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{X},$$

$$\mathcal{H}_{\boldsymbol{\beta}\sigma^2} = \frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \sigma^2} = -\frac{1}{\sigma^4} \mathbf{e}^\top \mathbf{X},$$

$$\mathcal{H}_{\sigma^2\boldsymbol{\beta}} = \frac{\partial^2 \ell}{\partial \sigma^2 \partial \boldsymbol{\beta}^\top} = -\frac{1}{\sigma^4} \mathbf{X}^\top \mathbf{e},$$

$$\mathcal{H}_{\sigma^2\sigma^2} = \frac{\partial^2 \ell}{(\partial \sigma^2)^2} = \frac{N}{2\sigma^4} - \frac{\mathbf{e}^\top \mathbf{e}}{\sigma^6}.$$

В точке истинных параметров  $\mathbf{e} = \boldsymbol{\varepsilon}$ . Используя это, получим, что компоненты информационной матрицы, вычисленной в точке истинных параметров равны:

$$\begin{aligned}\mathcal{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\theta}_0) &= E(\mathbf{g}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0)\mathbf{g}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0)^\top) = E\left(\frac{1}{\sigma^4} \mathbf{X}^\top \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top \mathbf{X}\right) = \\ &= \frac{1}{\sigma^4} \mathbf{X}^\top E(\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top) \mathbf{X} = \frac{1}{\sigma^4} \sigma^2 \mathbf{X}^\top \mathbf{I} \mathbf{X} = \frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{X},\end{aligned}$$

$$\mathcal{I}_{\boldsymbol{\beta}\sigma^2}(\boldsymbol{\theta}_0) = E(\mathbf{g}_{\sigma^2}(\boldsymbol{\theta}_0)\mathbf{g}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0)^\top) = E\left(\frac{1}{\sigma^6} (\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} - N\sigma^2) \boldsymbol{\varepsilon}^\top \mathbf{X}\right) = \mathbf{0}^\top,$$

$$\mathcal{I}_{\sigma^2\boldsymbol{\beta}}(\boldsymbol{\theta}_0) = \mathbf{0} \text{ (аналогично),}$$

$$\begin{aligned}\mathcal{I}_{\sigma^2\sigma^2}(\boldsymbol{\theta}_0) &= E(\mathbf{g}_{\sigma^2}(\boldsymbol{\theta}_0)^2) = E\left(\frac{1}{4\sigma^8} (\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} - N\sigma^2)^2\right) =, \\ &= \frac{1}{4\sigma^8} E((\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon})^2) - 2N \frac{1}{4\sigma^6} E(\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}) + N^2 \frac{1}{4\sigma^4} = \\ &= \frac{1}{4\sigma^8} (\sum_i E(\varepsilon_i^4) + \sum_{i \neq s} E(\varepsilon_i^2 \varepsilon_s^2)) - 2N \frac{1}{4\sigma^6} \sum_i E(\varepsilon_i^2) + N^2 \frac{1}{4\sigma^4} = \\ &= \frac{1}{4\sigma^8} 3N\sigma^4 + \frac{1}{4\sigma^8} (N^2 - N)\sigma^4 - 2N \frac{1}{4\sigma^6} N\sigma^2 + N^2 \frac{1}{4\sigma^4} = N \frac{1}{2\sigma^4}.\end{aligned}$$

В данном случае  $\mathcal{I}$  — блочно-диагональная матрица по параметрам  $\boldsymbol{\beta}$  и  $\sigma^2$ . В дальнейшем мы рассмотрим, какие полезные свойства из этого вытекают.

## Характеристика ММП

В статистике применяются три основных метода оценивания:

- Метод наименьших квадратов.
- (Обобщенный) метод моментов.
- Метод максимального правдоподобия.

Интересно сравнить ММП с двумя другими методами.

Условия, при которых можно использовать ММП более ограничительны. Метод требует явного задания вида распределения.

С другой стороны, ММП более универсален. Его можно использовать для любых моделей, задающих вид распределения наблюдаемых переменных. Два других метода можно использовать лишь тогда, когда распределение переменных можно представить в определенном виде. Если есть гипотеза о точном виде распределения, то всегда понятно, как получать оценки параметров, распределений параметров и различных статистик, как проверять гипотезы, хотя сами расчеты могут быть сложными.

Еще одно свойство — **инвариантность** по отношению к переобозначению параметров. Пусть  $\varphi(\cdot): \rho^k \rightarrow \rho^k$  однозначная обратимая функция. Можно подставить в функцию правдоподобия вместо  $\theta$  величину  $\varphi(\tau)$ , где  $\tau$  — новый вектор параметров,  $\tau \in \varphi^{-1}(\theta)$ . При этом, если  $\hat{\tau}$  — оценка МП в новой задаче, то  $\hat{\theta}$  — оценка МП в старой задаче.

Из инвариантности следует, что оценка МП как правило не может быть несмещенной. Пусть, например,  $E(\hat{\theta}) = \theta_0$ , где  $\theta_0$  — истинное значение параметра. Тогда оценка  $\hat{\tau}$ , полученная нелинейным преобразованием  $\theta = \varphi(\tau)$  будет смещенной:  $E(\hat{\tau}) \neq \tau_0$ , где  $\tau_0 = E(\varphi^{-1}(\hat{\theta}))$ .

Если правильно выбрать параметризацию, то распределение оценок в малых выборках может быть близко к асимптотическому, если неправильно, то асимптотическое распределение будет очень плохой аппроксимацией.

ММП получил широкое распространение благодаря своим хорошим асимптотическим свойствам:

- состоятельность,
- асимптотическая нормальность,
- асимптотическая эффективность.

С точки зрения эффективности сильные предположения о виде распределения, которые приходится делать, применяя ММП, окупаются (в большей

или меньшей степени). Поскольку мы делаем очень ограничительные предположения, то можем доказать более сильные утверждения.

## **Связь ММП с МНК. Квази-МП методы**

Хотя оценки МП являются специфическими по отношению к определенному виду распределения, значение метода может быть шире.

Идея состоит в том, чтобы процедуру получения оценок для одного распределения распространить на “близкие” распределения. Также методы получили название квази- или псевдо-ММП.

Метод максимального правдоподобия используют для нахождения способа расчетов, а затем уже доказывают, какими свойствами обладает этот метод по отношению к некоторому более широкому классу распределений.

Как мы видели, например, ММП в случае регрессии с нормально распределенными ошибками дает МНК, который на самом деле обладает “хорошими” свойствами и при ошибках, которые уже не имеют нормального распределения (хотя эффективность теряется).

Есть и обратная связь между этими двумя методами. МНК можно использовать как вычислительную процедуру, которая помогает находить оценки МП и строить тесты. Такое техническое использование МНК называют **вспомогательной регрессией**. Кроме того, вслед за Дэвидсоном и МакКинноном будем использовать термин **искусственная регрессия**, если вспомогательную регрессию можно применять как для нахождения оценок, так и для проверки гипотез относительно полученных оценок и проверки правильности спецификации модели.

## Связь гессиана и матрицы вкладов в градиент с информационной матрицей

### Гессиан и информационная матрица

Покажем, какая связь существует между информационной матрицей и гессианом. Сделаем это только в случае непрерывного распределения. Тот же метод доказательства очевидным образом распространяется на дискретные распределения. Применяя правило дифференцирования логарифма к логарифмической функции правдоподобия, получим следующее тождество:

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}} = \frac{1}{\mathcal{L}} \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}.$$

Продифференцируем по  $\boldsymbol{\theta}^\top$ :

$$\frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = \frac{1}{\mathcal{L}} \frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} - \frac{1}{\mathcal{L}^2} \frac{\mathcal{L}}{\partial \boldsymbol{\theta}^\top} \frac{\mathcal{L}}{\partial \boldsymbol{\theta}}.$$

Отсюда, опять воспользовавшись правилом дифференцирования логарифма, получим

$$\mathcal{H} = \frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = \frac{1}{\mathcal{L}} \frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} - \frac{\ell}{\partial \boldsymbol{\theta}^\top} \frac{\ell}{\partial \boldsymbol{\theta}}.$$

Найдем теперь ожидание обеих частей в точке  $\boldsymbol{\theta}_0$  (при истинных параметрах распределения):

$$\begin{aligned} \mathbb{E}(\mathcal{H}(Y, \boldsymbol{\theta}_0)) &= \mathbb{E}\left(\frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}(\boldsymbol{\theta}_0)\right) = \\ &= \int_{\mathcal{Y}} \mathcal{L}(\boldsymbol{\theta}_0, Y) \frac{1}{\mathcal{L}(\boldsymbol{\theta}_0, Y)} \frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}_0, Y)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} dY - \mathbb{E}\left(\frac{\ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^\top} \frac{\ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right). \end{aligned}$$

Второй член разности есть по определению информационная матрица  $\mathcal{I}(\boldsymbol{\theta}_0)$ . Интеграл равен нулю при условии, что операции интегрирования и дифференцирования перестановочны (для этого достаточно, в частности, чтобы область значения зависимой переменной  $\mathcal{Y}$  не зависела от  $\boldsymbol{\theta}$  или плотность распределения по краям  $\mathcal{Y}$  стремилась к нулю):

$$\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \int \mathcal{L}(\boldsymbol{\theta}_0, Y) dY = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} 1 = 0.$$

Таким образом, используя для краткости обозначения  $\mathcal{H}(Y, \boldsymbol{\theta}_0) = \mathcal{H}_0$  и  $\mathcal{I}(\boldsymbol{\theta}_0) = \mathcal{I}_0$ ,

$$- E(\mathcal{H}_0) = \mathcal{I}_0$$

— информационная матрица равна математическому ожиданию гессиана функции правдоподобия со знаком минус. То же самое свойство верно асимптотически (опять обозначаем  $\mathcal{I}^\infty(\boldsymbol{\theta}_0) = \mathcal{I}_0^\infty$ ):

$$- \lim_{N \rightarrow \infty} \frac{1}{N} E(\mathcal{H}_0) = \mathcal{I}_0^\infty.$$

### Матрица вкладов в градиент и информационная матрица

Прежде всего докажем, что математическое ожидание градиента в точке  $\boldsymbol{\theta}_0$  равно нулю ( $E \mathbf{g}(Y, \boldsymbol{\theta}_0) = 0$ ):<sup>2</sup>

$$\begin{aligned} E \mathbf{g}(Y, \boldsymbol{\theta}_0) &= \int \mathbf{g}(Y, \boldsymbol{\theta}_0) \mathcal{L}(Y, \boldsymbol{\theta}_0) dY = \int \frac{\partial \ell}{\partial \boldsymbol{\theta}}(Y, \boldsymbol{\theta}_0) \mathcal{L}(Y, \boldsymbol{\theta}_0) dY = \\ &= \int \frac{1}{\mathcal{L}(Y, \boldsymbol{\theta}_0)} \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}(Y, \boldsymbol{\theta}_0) \mathcal{L}(Y, \boldsymbol{\theta}_0) dY = \int \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}(Y, \boldsymbol{\theta}_0) dY = \\ &= \frac{\partial}{\partial \boldsymbol{\theta}} \int \mathcal{L}(Y, \boldsymbol{\theta}_0) dY = \frac{\partial}{\partial \boldsymbol{\theta}} 1 = 0. \end{aligned}$$

Как уже говорилось, функцию правдоподобия можно разбить по вкладам отдельных наблюдений:  $\ell(Y, \boldsymbol{\theta}) = \sum_i \ell_i(Y_i, \boldsymbol{\theta})$ . То же самое можно проделать с градиентом. Определим **матрицу вкладов в градиент** отдельных наблюдений  $\mathbf{G}$  как

$$G_{ij}(\boldsymbol{\theta}) = \frac{\partial \ell_i}{\partial \theta_j}(\boldsymbol{\theta}).$$

$$\text{При этом } \sum_i G_{ij} = \sum_i \frac{\partial \ell_i}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \sum_i \ell_i = \frac{\partial \ell}{\partial \theta_j} = g_j.$$

Используя рассуждения, аналогичные приведенным выше, можно показать, что  $E G_{ij}(Y, \boldsymbol{\theta}_0) = 0$ .

Мы так разделили функцию правдоподобия на вклады отдельных наблюдений, что  $E(\mathbf{G}_i(Y, \boldsymbol{\theta}_0) \mathbf{G}_s(Y, \boldsymbol{\theta}_0)^\top) = 0$ , где  $\mathbf{G}_i(Y, \boldsymbol{\theta}_0)$  и  $\mathbf{G}_s(Y, \boldsymbol{\theta}_0)$  — строки матрицы  $\mathbf{G}_0 = \mathbf{G}(Y, \boldsymbol{\theta}_0)$ , относящиеся к разным наблюдениям  $i$  и  $s$ . (Поскольку

<sup>2</sup> Мы опять предполагаем здесь, что операции интегрирования и дифференцирования перестановочны.)

элементы матрицы  $\mathbf{G}_0$  имеют нулевое математическое ожидание, то это означает что строки матрицы  $\mathbf{G}_0$ , относящиеся к разным наблюдениям, некоррелированы.) Докажем это свойство.

Функция правдоподобия  $i$ -го наблюдения по определению есть плотность распределения  $Y_i$  (в случае непрерывного распределения) условная по информации, содержащейся в наблюдениях  $1, \dots, i-1$  (условная по  $Y_1, \dots, Y_{i-1}$ ). Обозначим это информационное множество  $\Omega_i$ . Будем вычислять математическое ожидание по частям — сначала условное, а потом от него безусловное (правило полного мат. ожидания). Предположим, что  $i < s$ . Тогда

$$\begin{aligned} E(\mathbf{G}_i(\mathbf{Y}, \boldsymbol{\theta}_0) \mathbf{G}_s(\mathbf{Y}, \boldsymbol{\theta}_0)^\top) &= E(E(\mathbf{G}_i(\mathbf{Y}, \boldsymbol{\theta}_0) \mathbf{G}_s(\mathbf{Y}, \boldsymbol{\theta}_0)^\top | \Omega_i)) = \\ &= E(\mathbf{G}_i(\mathbf{Y}, \boldsymbol{\theta}_0) E(\mathbf{G}_s(\mathbf{Y}, \boldsymbol{\theta}_0)^\top | \Omega_i)) = 0. \end{aligned}$$

Равенство  $E(\mathbf{G}_s(\mathbf{Y}, \boldsymbol{\theta}_0)^\top | \Omega_i) = 0$  доказывается в точности по той же схеме, что и доказанное выше  $E \mathbf{g}(\mathbf{Y}, \boldsymbol{\theta}_0) = 0$ .

Используя это свойство, получим

$$E(\mathbf{G}_0^\top \mathbf{G}_0) = E\left(\sum_i \mathbf{G}_{0i}^\top \mathbf{G}_{0i}\right) = E\left(\left(\sum_i \mathbf{G}_{0i}\right)^\top \left(\sum_i \mathbf{G}_{0i}\right)\right) = E(\mathbf{g}_0 \mathbf{g}_0^\top).$$

Последнее выражение есть по определению информационная матрица. Таким образом,

$E(\mathbf{G}_0^\top \mathbf{G}_0) = \mathcal{I}_0.$
------------------------------------------------------

### Вычисление информационной матрицы

Рассмотрим теперь, как вычислить для конкретной модели информационную матрицу  $\mathcal{I}(\boldsymbol{\theta})$ . Здесь существуют три способа. Понятно, что все три способа должны для “хороших” моделей давать один и тот же результат. Во-первых, можно воспользоваться определением информационной матрицы:  $\mathcal{I} = E(\mathbf{g}\mathbf{g}^\top)$ . Во-вторых, можно воспользоваться равенством  $\mathcal{I}_0 = -E(\mathcal{H}_0)$ .

Самым простым часто (а именно тогда, когда функцию правдоподобия можно простым образом разбить на вклады наблюдений) оказывается третий способ, который использует только что рассмотренное свойство

$$\mathcal{I}_0 = E(\mathbf{G}_0^\top \mathbf{G}_0) = \sum_i E(\mathbf{G}_{i0}^\top \mathbf{G}_{i0}).$$

Выше была получено выражение для информационной матрицы в случае линейной регрессии с нормально распределенными ошибками прямо по определению. Вычислим теперь ее двумя другими способами.

Гессиан уже был вычислен выше. Математическое ожидание от него со знаком минус равно.

$$\mathcal{I}_0 = -E(\mathcal{H}_0) = -E \left( \begin{bmatrix} -\frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{X} & -\frac{1}{\sigma^4} \mathbf{X}^\top \boldsymbol{\varepsilon} \\ -\frac{1}{\sigma^4} \boldsymbol{\varepsilon}^\top \mathbf{X} & \frac{N}{2\sigma^4} - \frac{\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}}{\sigma^6} \end{bmatrix} \right) = \begin{bmatrix} \frac{\mathbf{X}^\top \mathbf{X}}{\sigma^2} & \mathbf{0} \\ \mathbf{0}^\top & \frac{N}{2\sigma^4} \end{bmatrix}.$$

Вклад в логарифмическую функцию правдоподобия  $i$ -го наблюдения равен

$$\ell_i = -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} (Y_i - \mathbf{X}_i \boldsymbol{\beta})^2.$$

Продифференцировав его, получим вклад в градиент  $i$ -го наблюдения в точке истинных параметров:

$$\mathbf{G}_{i0} = \left( \frac{1}{\sigma^2} \mathbf{X}_i^\top \boldsymbol{\varepsilon}_i, \frac{\boldsymbol{\varepsilon}_i^2}{2\sigma^4} - \frac{1}{2\sigma^2} \right).$$

Вклад в информационную матрицу  $i$ -го наблюдения в точке истинных параметров равен

$$\mathcal{I}_{i0} = E(\mathbf{G}_{i0}^\top \mathbf{G}_{i0}) = \begin{bmatrix} \frac{1}{\sigma^2} \mathbf{X}_i^\top \mathbf{X}_i & \mathbf{0} \\ \mathbf{0}^\top & \frac{1}{2\sigma^4} \end{bmatrix}.$$

Таким образом,

$$\mathcal{I}_0 = \sum_i \mathcal{I}_{i0} = \begin{bmatrix} \frac{\mathbf{X}^\top \mathbf{X}}{\sigma^2} & \mathbf{0} \\ \mathbf{0}^\top & \frac{N}{2\sigma^4} \end{bmatrix}.$$

Все три способа, как и следовало ожидать, привели к одному и тому же результату.

Заметим попутно, что  $\mathcal{I}_{i0}$  — положительно определенная матрица,  $\mathcal{I}_0$  при любом количестве наблюдений — положительно определенная матрица (в предположении, что матрица регрессоров имеет полный ранг). Из этого можно сделать вывод, что информация в некотором смысле увеличивается с ростом количества наблюдений. Это одно из объяснений названия "информационная матрица". В частности, определитель информационной матрицы увеличивается с ростом количества наблюдений:

$$|\mathcal{I}_0^{N+1}| > |\mathcal{I}_0^N|.$$

## Распределение градиента и оценок максимального правдоподобия

### Асимптотическое распределение градиента и оценок максимального правдоподобия

Оценки максимального правдоподобия имеют нормальное асимптотическое распределение. Для доказательства этого мы воспользуемся предположением, что градиент функции правдоподобия в точке истинных значений параметров  $\theta_0$  имеет асимптотическое нормальное распределение.

Градиент  $\mathbf{g}(\mathbf{Y}, \theta_0)$  будет иметь нормальное распределение (асимптотически), если к нему применима *центральная предельная теорема*. Надо представить  $\mathbf{g}_0$  как сумму некоторой последовательности случайных величин. Для этого подходит разложение градиента на вклады отдельных наблюдений

$$\mathbf{g}_i(\mathbf{Y}, \theta_0) = \sum_i G_{ij}(\mathbf{Y}, \theta_0).$$

Как сказано выше, каждое из слагаемых здесь имеет нулевое математическое ожидание. Если выполнены некоторые условия регулярности (см. литературу, посвященную центральной предельной теореме), то  $\sum G_{ij}(\mathbf{Y}, \theta_0)$  стремится к нормальному распределению с ростом количества наблюдений. Ковариационная матрица градиента в точке  $\theta_0$  есть информационная матрица, поскольку его математическое ожидание равно нулю:  $V(\mathbf{g}_0) = E(\mathbf{g}_0 \mathbf{g}_0^T) = \mathcal{I}$ . Последнее равенство выполнено по определению.

Окончательно получаем

$$\frac{1}{\sqrt{N}} \mathbf{g}_0 \stackrel{a}{\sim} N(\mathbf{0}, \mathcal{I}_0^\infty).$$

Используя это свойство градиента мы докажем асимптотическую нормальность оценок ММП. Для этого используем разложение в ряд Тейлора в точке  $\theta_0$  до членов первого порядка:

$$\mathbf{0} = \mathbf{g}(\hat{\theta}) = \mathbf{g}(\theta_0) + \mathcal{H}(\bar{\theta})(\hat{\theta} - \theta_0),$$

где  $\mathcal{H}$  — гессиан (матрица вторых производных от логарифмической функции правдоподобия),  $\bar{\theta}_j$  — выпуклая комбинация  $\hat{\theta}_j$  и  $\theta_{0j}$ . Поскольку  $\hat{\theta}_j$  — состоятельная оценка параметра  $\theta_{0j}$ , то  $\bar{\theta}_j$  тоже должна быть состоятельной оценкой  $\theta_{0j}$ . Поскольку  $-\frac{1}{N} \mathcal{H}_0 \stackrel{a}{=} \mathcal{I}_0^\infty$ , то имеем асимптотическое равенство:  $-\frac{1}{N} \mathcal{H}(\bar{\theta}) \stackrel{a}{=} \mathcal{I}_0^\infty$ .

Таким образом,  $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{\sim} (\mathcal{I}_0^\infty)^{-1} \frac{1}{\sqrt{N}} \mathbf{g}_0 \stackrel{a}{\sim} \mathbf{N}(\mathbf{0}, (\mathcal{I}_0^\infty)^{-1} \mathcal{I}_0^\infty (\mathcal{I}_0^\infty)^{-1})$ .

Окончательно получим

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{\sim} \mathbf{N}(\mathbf{0}, (\mathcal{I}_0^\infty)^{-1}).$$

Это соотношение позволяет оценить ковариационную матрицу оценок  $\hat{\boldsymbol{\theta}}$ . С этой точки зрения оценка обратной информационной матрицы является оценкой ковариационной матрицы МП-оценок  $\hat{\mathcal{V}}_{\boldsymbol{\theta}}$  (с точностью до множителя  $1/N$ ), и эти термины можно использовать как синонимы. Понятно, что для этого должны быть выполнены соответствующие условия, гарантирующие, что операции интегрирования и дифференцирования коммутируют и что применима центральная предельная теорема, что мы всегда в дальнейшем будем предполагать.

### **Выборочная оценка распределения градиента и оценок максимального правдоподобия**

Для получения выборочной оценки распределений МП-оценок  $\hat{\boldsymbol{\theta}}$  и градиента, можно воспользоваться формулами для их асимптотических распределений. Все эти величины асимптотически нормально распределены и их асимптотические матрицы ковариаций являются функциями асимптотической информационной матрицы в точке истинных параметров  $(\mathcal{I}_0^\infty)$ . Таким образом, требуется получить состоятельную оценку  $\mathcal{I}_0^\infty$ , чтобы подставить ее в соответствующие формулы. Будем обозначать символом  $\hat{\mathcal{I}}$  такую матрицу, что  $1/N \hat{\mathcal{I}}$  — состоятельная оценка  $\mathcal{I}_0^\infty$ :

$$\text{Plim}_{N \rightarrow \infty} \frac{1}{N} \hat{\mathcal{I}} = \mathcal{I}_0^\infty.$$

Поскольку  $\hat{\boldsymbol{\theta}}$  — состоятельная оценка истинных параметров  $\boldsymbol{\theta}_0$ , то  $1/N \mathcal{I}(\hat{\boldsymbol{\theta}})$  — состоятельная оценка  $\mathcal{I}_0^\infty$  (по теореме Слуцкого). Это дает первый способ оценивания. Он состоит в том, чтобы сначала для данной модели найти функцию  $\mathcal{I}(\boldsymbol{\theta})$ , а затем подставить в нее оценки максимального правдоподобия  $\hat{\boldsymbol{\theta}}$  (конечно, подойдут и любые другие состоятельные оценки). Методы нахождения  $\mathcal{I}(\boldsymbol{\theta})$  описаны ниже.

Другой способ основывается на равенстве для информационной матрицы  $\mathcal{I}_0^\infty = -\lim_{N \rightarrow \infty} 1/N \mathbf{E}(\mathcal{H}_0)$  и на том, что ожидаемый гессиан  $\mathbf{E}(\mathcal{H}_0)$  асимп-

тотически равен **эмпирическому гессиану**  $\hat{\mathcal{H}} = \mathcal{H}(Y, \hat{\theta}) = \frac{\partial^2 \ell}{\partial \theta \partial \theta^\top}(Y, \hat{\theta})$ . Этот способ обычно проще предыдущего, поскольку не требует вычисления математических ожиданий. Получить матрицу вторых производных данной функции правдоподобия можно и с помощью компьютерной программы.

Особой простотой, и потому притягательностью (требуется найти только первые производные), отличается третий способ оценивания информационной матрицы, использующий матрицу вкладов в градиент  $G$ . Этот способ предложен в статье Berndt, Hall, Hall, and Hausman (1974) и поэтому называется ВННН. Другое название — метод внешнего произведения градиента (outer product of the gradient, сокращенно OPG). Этот способ основан на том, что  $E(G_0^\top G_0) = \mathcal{I}_0$ . Предлагается использовать матрицу  $G(Y, \hat{\theta})^\top G(Y, \hat{\theta})$  в качестве  $\hat{\mathcal{I}}$ .

Таким образом, имеем три варианта матрицы  $\hat{\mathcal{I}}$ :

$$\text{I. } \mathcal{I}(\hat{\theta}); \quad \text{II. } \mathcal{H}(Y, \hat{\theta}); \quad \text{III. } G(Y, \hat{\theta})^\top G(Y, \hat{\theta}).$$

Как показывают эксперименты методом Монте-Карло,  $G(Y, \hat{\theta})^\top G(Y, \hat{\theta})$  — самая неточная оценка  $\hat{\mathcal{I}}$  в конечных выборках, а тесты, основанные на  $\mathcal{I}(\hat{\theta})$  обычно не уступают тестам, основанным на  $\mathcal{H}(Y, \hat{\theta})$ .

Три рассмотренных способа нахождения  $\hat{\mathcal{I}}$  подходят для любых распределений. Есть также более специфические методы, которые можно использовать только в случае моделей определенного вида. Например, метод Гаусса-Ньютона используется в нелинейных регрессиях, метод удвоенной регрессии — в квазирегрессионных моделях с неизвестными параметрами в правой части.

Особого рассмотрения требует нахождение оценки ковариационной матрицы оценок в случае квази-МП методов (их называют также псевдо-МП методами). Если предполагается, что ошибки в модели имеют нормальное распределение и гомоскедастичны, а на самом деле это не так, то часто только что рассмотренные методы дают несостоятельные оценки. Оказывается, что во многих случаях следующие оценки состоятельны (конечно, при вычислении этих величин используется не настоящая, а псевдо функция правдоподобия):

$$\begin{aligned} & \mathcal{H}(Y, \hat{\theta})^{-1} \mathcal{I}(\hat{\theta}) \mathcal{H}(Y, \hat{\theta})^{-1}. \\ & \mathcal{H}(Y, \hat{\theta})^{-1} G(Y, \hat{\theta})^\top G(Y, \hat{\theta}) \mathcal{H}(Y, \hat{\theta})^{-1}. \end{aligned}$$

Поясним интуитивно, откуда берутся эти формулы. При выводе асимптотического распределения оценок максимального правдоподобия, мы пользовались тем, что “усредненный” гессиан  $-1/N \mathcal{H}_0$  равен асимптотически  $\mathcal{I}_0^\infty$ . В общем случае нужно воспользоваться пределом  $1/N E(\mathcal{H})$  — “асимптотическим” ожидаемым гессианом в точке истинных оценок  $(\mathcal{H}_0^\infty)$ . Формула приобретет следующий вид:

$$\sqrt{N} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{=} (\mathcal{H}_0^\infty)^{-1} \frac{1}{\sqrt{N}} \mathbf{g}_0 \stackrel{a}{\sim} N(\mathbf{0}, (\mathcal{H}_0^\infty)^{-1} \mathcal{I}_0^\infty (\mathcal{H}_0^\infty)^{-1}).$$

## Численные методы нахождения оценок максимального правдоподобия

Рассмотрим семейство универсальных алгоритмов вычисления оценок максимального правдоподобия, тесно связанных с только что рассмотренными способами получения матрицы  $\hat{\mathcal{I}}$ . Эти алгоритмы являются итеративными градиентными методами и  $t$ -й шаг алгоритма задается формулой

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + (\hat{\mathcal{I}}^t)^{-1} \mathbf{g}(\boldsymbol{\theta}^t).$$

Стационарная точка этого процесса  $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t$  будет удовлетворять уравнениям правдоподобия  $\mathbf{g}=\mathbf{0}$  и (с соответствующими оговорками) будет оценкой максимального правдоподобия.

Если в качестве  $\hat{\mathcal{I}}^t$  взять информационную матрицу в точке оценок  $\mathcal{I}(\boldsymbol{\theta}^t)$ , то мы получаем метод, называемый по-английски **method of scoring**:

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \mathcal{I}(\boldsymbol{\theta}^t)^{-1} \mathbf{g}(\boldsymbol{\theta}^t).$$

Если в качестве  $\hat{\mathcal{I}}^t$  взять минус гессиан  $-\mathcal{H}(\boldsymbol{\theta}^t)$ , то мы получаем классический **метод Ньютона**:

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \mathcal{H}(\boldsymbol{\theta}^t)^{-1} \mathbf{g}(\boldsymbol{\theta}^t).$$

Метод Ньютона, как правило, быстрее сходится в ближайшей окрестности оценок МП, зато метод, использующий информационную матрицу обычно менее чувствителен к выбору начальных приближений.

Шаг метода ВННН (OPG) можно получить с помощью вспомогательной (искусственной) регрессии, зависимой переменной в которой будет вектор, составленный из единиц (обозначим его  $\mathbf{1}$ ), а матрицей регрессоров — матрица  $\mathbf{G}(\boldsymbol{\theta}^t)$ . Если  $\Delta\boldsymbol{\theta}^t$  — оценки коэффициентов в этой вспомогательной регрессии на  $t$ -м шаге, то итерация имеет вид

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \Delta\boldsymbol{\theta}^t, \quad \text{где } \Delta\boldsymbol{\theta}^t = (\mathbf{G}(\boldsymbol{\theta}^t)^\top \mathbf{G}(\boldsymbol{\theta}^t))^{-1} \mathbf{G}(\boldsymbol{\theta}^t)^\top \mathbf{1}.$$

Хотя этот последний алгоритм является самым простым, но, как правило, сходится очень медленно. Если учесть, что обычно при использовании этого метода  $(\mathbf{G}(\boldsymbol{\theta}^t)^\top \mathbf{G}(\boldsymbol{\theta}^t))^{-1}$  берут в качестве оценки ковариационной матрицы оценок, то использовать его нежелательно.

Возможны различные модификации этой основной идеи.

Шаг алгоритма можно вычислять, домножая исходный шаг на параметр  $\lambda$ :

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \lambda (\hat{\mathcal{I}}^t)^{-1} \mathbf{g}(\boldsymbol{\theta}^t).$$

Разумно выбрать параметр  $\lambda$ , максимизируя по нему функцию правдоподобия в точке  $\boldsymbol{\theta}^{t+1}$ :

$$\lambda = \operatorname{argmax} \lambda(\boldsymbol{\theta}^t + \lambda (\hat{\mathcal{I}}^t)^{-1} \mathbf{g}(\boldsymbol{\theta}^t)).$$

В частном случае матрица  $\mathcal{I}_0$  является блочно-диагональной. Тогда шаг алгоритма можно разбить на несколько “подшагов”, один для каждого блока. Изменяются при этом только параметры, соответствующие данному блоку.

Если из условий первого порядка выразить одни оцениваемые параметры через другие и подставить их в функцию правдоподобия, то получится **концентрированная функция правдоподобия**. Действуя таким образом, задачу поиска оценок МП можно упростить, сведя к задаче максимизации концентрированной функции правдоподобия по меньшему числу параметров. Задача может упроститься до одномерного поиска.

Существует много других алгоритмов. Есть алгоритмы специально сконструированные для конкретной модели; с примерами их мы встретимся в дальнейшем. Есть универсальные методы, которые можно применять к широкому классу моделей, такие как метод удвоенной регрессии и итеративный обобщенный МНК. Можно, конечно, использовать универсальные оптимизационные алгоритмы, которые подходят не только для максимизации функции правдоподобия.

## ММП и проверка гипотез

### Асимптотическое распределение и асимптотическая эквивалентность трех классических статистик

Предположим, что мы хотим проверить гипотезу о том, что вектор истинных параметров  $\theta_0$  удовлетворяет набору ограничений, который в векторном виде можно записать как

$$r(\theta_0) = 0.$$

Тогда с учетом этой информации задача получения оценки максимального правдоподобия эквивалентна задаче нахождения седловой точки лагранжиана:

$$L(\theta, \lambda) = \ell(\theta) - r^T(\theta) \lambda.$$

Ограниченная оценка  $\tilde{\theta}$  должна вместе с вектором множителей Лагранжа  $\tilde{\lambda}$  удовлетворять следующей системе условий первого порядка:

$$\begin{aligned} g(\tilde{\theta}) &= R^T(\tilde{\theta}) \tilde{\lambda}, \\ r(\tilde{\theta}) &= 0, \end{aligned}$$

где  $R(\theta)$  — матрица первых производных ограничений:  $R = \frac{\partial r}{\partial \theta}$ .

Для вывода распределений интересующих нас статистик используем тот же прием, с помощью которого выше получено распределение оценок. Поскольку мы предполагаем, что оценки МП состоятельны и нас интересуют асимптотические распределение, то для разложений в ряд Тейлора будем писать приближенные равенства. Более строгие рассуждения должны быть аналогичны использованным выше.

Разложим градиент и ограничения в ряд Тейлора до членов первого порядка в точке истинных параметров  $\theta_0$ :

$$\begin{aligned} g(\tilde{\theta}) &\approx g_0 + \mathcal{H}_0(\tilde{\theta} - \theta_0), \\ r(\tilde{\theta}) &\approx R_0(\tilde{\theta} - \theta_0). \end{aligned}$$

При получении второго соотношения мы использовали, что в точке истинных параметров ограничения выполняются:  $r(\theta_0) = 0$ .

Подставив эти приближения в условия первого порядка, получим следующие асимптотические равенства:

$$\begin{aligned} g_0 + \mathcal{H}_0(\tilde{\theta} - \theta_0) &\stackrel{a}{=} R_0^T \tilde{\lambda}, \\ R_0(\tilde{\theta} - \theta_0) &\stackrel{a}{=} 0. \end{aligned}$$

Перепишем систему в блочной форме:

$$\begin{bmatrix} -\mathcal{H}_0 & \mathbf{R}_0^\top \\ \mathbf{R}_0 & \mathbf{0} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \\ \tilde{\boldsymbol{\lambda}} \end{bmatrix} \stackrel{a}{=} \begin{bmatrix} \mathbf{g}_0 \\ \mathbf{0} \end{bmatrix}.$$

Отсюда, домножая на  $\sqrt{N}$  и  $\frac{1}{\sqrt{N}}$ , чтобы получились величины порядка  $O(1)$ , получим асимптотическое равенство:

$$\begin{bmatrix} \sqrt{N}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ \frac{1}{\sqrt{N}}\tilde{\boldsymbol{\lambda}} \end{bmatrix} \stackrel{a}{=} \begin{bmatrix} \mathcal{I}_0^\infty & \mathbf{R}_0^\top \\ \mathbf{R}_0 & \mathbf{0} \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{\sqrt{N}}\mathbf{g}_0 \\ \mathbf{0} \end{bmatrix}.$$

Используем следующее правило блочного обращения матрицы:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}\mathbf{A}^{-1}\mathbf{B} - \mathbf{D})^{-1}\mathbf{C}\mathbf{A}^{-1} & \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}\mathbf{A}^{-1}\mathbf{B} - \mathbf{D})^{-1} \\ (\mathbf{C}\mathbf{A}^{-1}\mathbf{B} - \mathbf{D})^{-1}\mathbf{C}\mathbf{A}^{-1} & -(\mathbf{C}\mathbf{A}^{-1}\mathbf{B} - \mathbf{D})^{-1} \end{bmatrix}.$$

В данном случае

$$\mathbf{A} = \mathcal{I}_0^\infty, \quad \mathbf{B} = \mathbf{R}_0^\top, \quad \mathbf{C} = \mathbf{R}_0, \quad \mathbf{D} = \mathbf{0}, \quad (\mathbf{C}\mathbf{A}^{-1}\mathbf{B} - \mathbf{D})^{-1} = (\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1}\mathbf{R}_0^\top)^{-1}.$$

Таким образом,

$$\begin{aligned} & \begin{bmatrix} \mathcal{I}_0^\infty & \mathbf{R}_0^\top \\ \mathbf{R}_0 & \mathbf{0} \end{bmatrix}^{-1} = \\ & = \begin{bmatrix} (\mathcal{I}_0^\infty)^{-1}(\mathbf{I} - \mathbf{R}_0^\top(\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1}\mathbf{R}_0^\top)^{-1}\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1}) & (\mathcal{I}_0^\infty)^{-1}\mathbf{R}_0^\top(\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1}\mathbf{R}_0^\top)^{-1} \\ (\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1}\mathbf{R}_0^\top)^{-1}\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1} & -(\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1}\mathbf{R}_0^\top)^{-1} \end{bmatrix}. \end{aligned}$$

Получим выражения, асимптотически эквивалентные оценкам  $\tilde{\boldsymbol{\theta}}$  и множителям Лагранжа  $\tilde{\boldsymbol{\lambda}}$ :

$$\sqrt{N}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{=} (\mathcal{I}_0^\infty)^{-1}(\mathbf{I} - \mathbf{R}_0^\top(\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1}\mathbf{R}_0^\top)^{-1}\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1})\left(\frac{1}{\sqrt{N}}\mathbf{g}_0\right), \quad (1)$$

$$\frac{1}{\sqrt{N}}\tilde{\boldsymbol{\lambda}} \stackrel{a}{=} (\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1}\mathbf{R}_0^\top)^{-1}\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1}\left(\frac{1}{\sqrt{N}}\mathbf{g}_0\right). \quad (2)$$

Вспомним, что  $\frac{1}{\sqrt{N}}\mathbf{g}_0 \stackrel{a}{\sim} \mathbf{N}(\mathbf{0}, \mathcal{I}_0^\infty)$ .

Отсюда получим асимптотическое распределение вектора множителей Лагранжа:

$$\frac{1}{\sqrt{N}}\tilde{\boldsymbol{\lambda}} \stackrel{a}{\sim} \mathbf{N}(\mathbf{0}, (\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1}\mathbf{R}_0^\top)^{-1}\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1}\mathcal{I}_0^\infty(\mathcal{I}_0^\infty)^{-1}\mathbf{R}_0^\top(\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1}\mathbf{R}_0^\top)^{-1}),$$

$$\frac{1}{\sqrt{N}}\tilde{\boldsymbol{\lambda}} \stackrel{a}{\sim} \mathbf{N}(\mathbf{0}, (\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1}\mathbf{R}_0^\top)^{-1}).$$

**Статистикой множителя Лагранжа** называют следующую величину:

$$\text{LM} \equiv \tilde{\boldsymbol{\lambda}}^{\top} \tilde{\mathbf{R}} \tilde{\mathcal{I}}^{-1} \tilde{\mathbf{R}}^{\top} \tilde{\boldsymbol{\lambda}}.$$

Здесь  $\tilde{\mathcal{I}}$  — матрица, полученная на основании выборочной информации в точке  $\tilde{\boldsymbol{\theta}}$ , такая что  $1/N \tilde{\mathcal{I}}$  — состоятельная оценка  $\mathcal{I}_0^{\infty}$ . Величина LM имеет распределение  $\chi^2$  с  $p$  степенями свободы, где  $p$  — размерность вектора ограничений  $\mathbf{r}$ :

$$\text{LM} \stackrel{a}{\sim} \chi^2(p).$$

Это следует из формулы для распределения  $\tilde{\boldsymbol{\lambda}}$ , состоятельности оценки  $\tilde{\boldsymbol{\theta}}$  и невырожденности матрицы  $\mathbf{R}_0 (\mathcal{I}_0^{\infty})^{-1} \mathbf{R}_0^{\top}$ .

Вспомним, что одно из условий первого порядка максимума функции правдоподобия имеет вид  $\tilde{\mathbf{g}} = \tilde{\mathbf{R}}^{\top} \tilde{\boldsymbol{\lambda}}$ . Это позволяет выразить статистику множителя Лагранжа через градиент логарифмической функции правдоподобия:

$$\text{LM} = \tilde{\mathbf{g}}^{\top} \tilde{\mathcal{I}}^{-1} \tilde{\mathbf{g}} \stackrel{a}{\sim} \chi^2(p).$$

Хотя статистика множителя Лагранжа получила свое название благодаря тому, что ее можно выразить через множители Лагранжа, на практике гораздо чаще используют градиентную форму (score form of LM test).

Если вспомнить асимптотическое выражение (2) для  $\tilde{\boldsymbol{\lambda}}$ , то можно выразить (асимптотически) LM-тест через  $\mathbf{g}_0$ :

$$\text{LM} \stackrel{a}{=} 1/N \mathbf{g}_0^{\top} (\mathcal{I}_0^{\infty})^{-1} \mathbf{R}_0^{\top} (\mathbf{R}_0 (\mathcal{I}_0^{\infty})^{-1} \mathbf{R}_0^{\top})^{-1} \mathbf{R}_0 (\mathcal{I}_0^{\infty})^{-1} \mathbf{g}_0.$$

**Статистика отношения правдоподобия** по определению есть

$$\text{LR} \equiv 2(\tilde{\ell} - \hat{\ell}).$$

Найдем ее асимптотическое распределение. Используем для этого разложение в ряд Тейлора:

$$\tilde{\ell} = \hat{\ell} + \frac{1}{2} (\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})^{\top} \mathcal{H}(\bar{\boldsymbol{\theta}}) (\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}),$$

где  $\bar{\boldsymbol{\theta}}_j$  — выпуклая линейная комбинация  $\tilde{\boldsymbol{\theta}}_j$  и  $\hat{\boldsymbol{\theta}}_j$ . Поскольку  $\tilde{\boldsymbol{\theta}}$  и  $\hat{\boldsymbol{\theta}}$  — состоятельные оценки  $\boldsymbol{\theta}_0$ , то  $-\mathcal{H}(\bar{\boldsymbol{\theta}}) \stackrel{a}{=} N \mathcal{I}_0^{\infty}$ .

$$\text{LR} = 2(\tilde{\ell} - \hat{\ell}) \stackrel{a}{=} N (\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})^{\top} \mathcal{I}_0^{\infty} (\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}).$$

Асимптотический эквивалент этой статистики также можно записать в терминах  $\boldsymbol{\theta}_0$ ,  $\mathcal{I}_0^\infty$ ,  $\mathbf{R}_0$  и  $\mathbf{g}_0$ .

Отняв от  $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{=} (\mathcal{I}_0^\infty)^{-1} \frac{1}{\sqrt{N}} \mathbf{g}_0$  доказанное ранее равенство (1) получаем, что  $\sqrt{N}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) \stackrel{a}{=} (\mathcal{I}_0^\infty)^{-1} \mathbf{R}_0^\top (\mathbf{R}_0 (\mathcal{I}_0^\infty)^{-1} \mathbf{R}_0^\top)^{-1} \mathbf{R}_0 (\mathcal{I}_0^\infty)^{-1} \mathbf{g}_0$ .

Отсюда следует, что статистика отношения правдоподобия асимптотически равна той же самой случайной величине, что и статистика множителя Лагранжа:

$$LR \stackrel{a}{=} 1/N \mathbf{g}_0^\top (\mathcal{I}_0^\infty)^{-1} \mathbf{R}_0^\top (\mathbf{R}_0 (\mathcal{I}_0^\infty)^{-1} \mathbf{R}_0^\top)^{-1} \mathbf{R}_0 (\mathcal{I}_0^\infty)^{-1} \mathbf{g}_0$$

Эта это означает, что статистика отношения правдоподобия также имеет асимптотическое распределение  $\chi^2$  с  $p$  степенями свободы.

Третья классическая статистика основана на распределении  $\mathbf{r}(\hat{\boldsymbol{\theta}})$ . Поскольку  $\hat{\boldsymbol{\theta}}$  — оценка, полученная без учета ограничений, то в общем случае  $\mathbf{r}(\hat{\boldsymbol{\theta}}) \neq 0$ , однако, если верна нулевая гипотеза, то  $\mathbf{r}(\hat{\boldsymbol{\theta}}) \stackrel{a}{=} 0$ . Разложим  $\hat{\mathbf{r}} = \mathbf{r}(\hat{\boldsymbol{\theta}})$  в ряд Тейлора в точке  $\boldsymbol{\theta}_0$ , учитывая, что  $\mathbf{r}(\boldsymbol{\theta}_0) = 0$ :

$$\hat{\mathbf{r}} \approx \mathbf{r}(\boldsymbol{\theta}_0) + \mathbf{R}_0(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \mathbf{R}_0(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

Ранее было выведено, что  $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{\sim} N(\mathbf{0}, \mathcal{I}_0)$ . Отсюда получим **статистику Вальда**:

$$W \equiv \hat{\mathbf{r}}^\top (\hat{\mathbf{R}} \hat{\mathcal{I}}^{-1} \hat{\mathbf{R}}^\top)^{-1} \hat{\mathbf{r}}$$

Здесь  $\hat{\mathcal{I}}$  — матрица, полученная на основании выборочной информации в точке  $\hat{\boldsymbol{\theta}}$ , такая что  $1/N \hat{\mathcal{I}}$  — состоятельная оценка  $\mathcal{I}_0^\infty$ .

Как и в случае двух других тестов  $W \stackrel{a}{\sim} \chi^2(p)$ .

В пределе  $\hat{\mathbf{r}} \stackrel{a}{=} \mathbf{R}_0(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{=} 1/N \mathbf{R}_0 (\mathcal{I}_0^\infty)^{-1} \mathbf{g}_0$ . Значит,

$$W \stackrel{a}{=} 1/N \mathbf{g}_0^\top (\mathcal{I}_0^\infty)^{-1} \mathbf{R}_0^\top (\mathbf{R}_0 (\mathcal{I}_0^\infty)^{-1} \mathbf{R}_0^\top)^{-1} \mathbf{R}_0 (\mathcal{I}_0^\infty)^{-1} \mathbf{g}_0.$$

Тем самым мы показали, что с ростом количества наблюдений все три статистики стремятся к одной и той же случайной переменной, которая имеет распределение  $\chi^2(p)$ . Другими словами, три классических теста **асимптотически эквивалентны**.

Все три статистики совпадают, если логарифмическая функция правдоподобия является квадратичной. Это верно, в частности, для линейной регрессии с известной дисперсией, например,

$$Y = X\beta + \epsilon, \text{ где } \epsilon_i \sim \text{NID}(0,1).$$

Рассмотрим, к примеру, логарифмическую функцию правдоподобия с единственным (скалярным) параметром  $\theta$ :

$$\ell = -(a - b\theta)^2 + \text{const.}$$

Гессиан  $\mathcal{H} = \frac{\partial^2 \ell}{\partial \theta^2} = -2b^2$  является постоянной величиной, информационная матрица, таким образом, равна  $\mathcal{I} = 2b^2$  при всех  $\theta$ .

Возьмем ограничение вида  $r(\theta) = 3\theta - 1$

Получим  $\hat{\theta} = \frac{a}{b}$ ,  $\tilde{\theta} = \frac{1}{3}$ . Откуда  $\hat{\ell} = 0 + \text{const}$ ,  $\tilde{\ell} = -(a - \frac{b}{3})^2 + \text{const}$ .

$$\text{LR} = 2(\hat{\ell} - \tilde{\ell}) = 2(a - \frac{b}{3})^2.$$

Градиент равен  $g = 2b(a - b\theta)$ . Таким образом,  $\tilde{g} = 2b(a - \frac{1}{3}b)$ .

$$\text{LM} = \tilde{g}^T \tilde{\mathcal{I}}^{-1} \tilde{g} = 2b(a - \frac{1}{3}b) \frac{1}{2b^2} 2b(a - \frac{1}{3}b) = 2(a - \frac{b}{3})^2.$$

Найдем ту же статистику через множитель Лагранжа.

$$L = -(a - b\theta)^2 - \lambda(3\theta - 1) \rightarrow \max_{\theta}$$

$$\frac{\partial L}{\partial \theta} = 2(a - b\theta) - 3\lambda = 0 \Rightarrow \tilde{\lambda} = \frac{2}{3}(a - b\tilde{\theta}) = \frac{2}{3}(a - \frac{b}{3}).$$

$$R = \frac{\partial r}{\partial \theta} = 3, \tilde{R} = 3.$$

$$\text{LM} = \tilde{\lambda}^T \tilde{R} \tilde{\mathcal{I}}^{-1} \tilde{R}^T \tilde{\lambda} = \frac{2}{3}(a - \frac{b}{3}) 3 \frac{1}{2b^2} 3 \frac{2}{3}(a - \frac{b}{3}) = 2(a - \frac{b}{3})^2.$$

Теперь найдем статистику Вальда.

$$\hat{r} = 3\hat{\theta} - 1 = 3\frac{a}{b} - 1. \quad \hat{R} = 3.$$

$$W = \hat{r}^T (\hat{R} \hat{\mathcal{I}}^{-1} \hat{R}^T)^{-1} \hat{r} = (3\frac{a}{b} - 1) (3 \frac{1}{2b^2} 3)^{-1} (3\frac{a}{b} - 1) = 2(a - \frac{b}{3})^2.$$

## Соотношения между статистиками

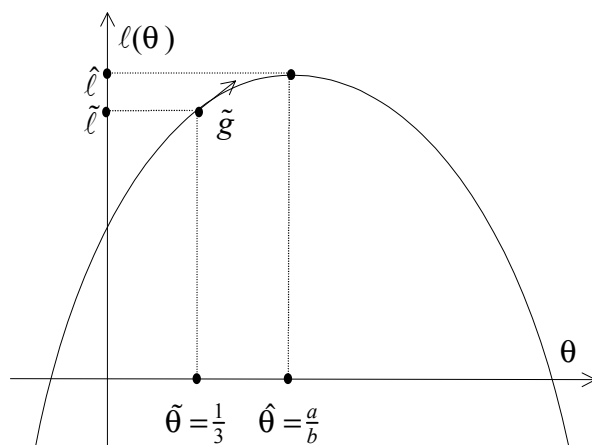


Рис. 1

Все три классические статистики совпадают в случае “бесконечно большой выборки”. В выборках конечных размеров их поведение может существенно отличаться от асимптотического. Поэтому не всегда на эти классические статистики можно полагаться. В этом их отличие от F-статистик, которые имеют *точное* распределение в конечной выборке в случае классической линейной регрессии с нормальными ошибками и линейной проверяемой гипотезой. Рассмотрим этот случай более подробно.

Предположим, что проверяется гипотеза

$$\mathbf{Q}\boldsymbol{\beta} = \mathbf{q},$$

где  $\mathbf{Q}$  — известная матрица ( $p \times m$ ),  $\mathbf{q}$  — известный вектор ( $p \times 1$ ). В использованных выше обозначениях  $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{r}(\boldsymbol{\beta}, \sigma^2) = \mathbf{Q}\boldsymbol{\beta} - \mathbf{q}$ , матрица  $\mathbf{R}(\boldsymbol{\theta})$  равна  $[\mathbf{Q} \ \mathbf{0}]$  при всех значениях  $\boldsymbol{\theta}$  (нулевой вектор относится к параметру  $\sigma^2$ ). Проверяемой гипотезе соответствует следующая статистика Вальда (вспомним, что  $\mathcal{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1} = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$ ):

$$\begin{aligned} W &= \hat{\mathbf{r}}^T (\hat{\mathbf{R}} \hat{\mathcal{I}}^{-1} \hat{\mathbf{R}}^T)^{-1} \hat{\mathbf{r}} = (\mathbf{Q}\hat{\boldsymbol{\beta}} - \mathbf{q})^T (\mathbf{Q} \hat{\mathcal{I}}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1} \mathbf{Q}^T)^{-1} (\mathbf{Q}\hat{\boldsymbol{\beta}} - \mathbf{q}) = \\ &= \frac{1}{\hat{\sigma}^2} (\mathbf{Q}\hat{\boldsymbol{\beta}} - \mathbf{q})^T (\mathbf{Q}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{Q}^T)^{-1} (\mathbf{Q}\hat{\boldsymbol{\beta}} - \mathbf{q}). \end{aligned}$$

Нам нужно максимизировать функцию правдоподобия при ограничении  $\mathbf{Q}\boldsymbol{\beta} = \mathbf{q}$ . Лагранжиан рассматриваемой задачи условной максимизации имеет вид

$$L = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) - (\mathbf{Q}\boldsymbol{\beta} - \mathbf{q})^T \boldsymbol{\lambda}.$$

В максимуме должно выполняться

$$\frac{\partial L}{\partial \beta^T}(\tilde{\theta}, \tilde{\lambda}) = \frac{1}{\tilde{\sigma}^2} X^T(Y - X\beta) - Q^T \tilde{\lambda} = 0.$$

Отсюда

$$\beta = (X^T X)^{-1} X^T Y - \tilde{\sigma}^2 (X^T X)^{-1} Q^T \tilde{\lambda} = \hat{\beta} - \tilde{\sigma}^2 (X^T X)^{-1} Q^T \tilde{\lambda},$$

где  $\hat{\beta} = (X^T X)^{-1} X^T Y$  — оценки ОМНК (без ограничений). Домножая это равенство слева на  $Q$  и учитывая, что  $Q\hat{\beta} = q$ , получим

$$\tilde{\sigma}^2 \tilde{\lambda} = (Q(X^T X)^{-1} Q^T)^{-1} (Q\hat{\beta} - q).$$

Таким образом, оценки с учетом ограничений равны

$$\tilde{\beta} = \hat{\beta} - (X^T X)^{-1} Q^T (Q(X^T X)^{-1} Q^T)^{-1} (Q\hat{\beta} - q).$$

Из условия  $\frac{\partial L}{\partial \beta^T}(\tilde{\theta}, \tilde{\lambda}) = 0$  несложно получить, что  $\tilde{\sigma}^2 = \frac{R\tilde{S}S}{N}$ , где  $R\tilde{S}S$  —

сумма квадратов остатков в регрессии с ограничениями (так же как  $\hat{\sigma}^2 = \frac{R\hat{S}S}{N}$  в регрессии без ограничений).

Статистика множителя Лагранжа равна:

$$\begin{aligned} LM &= \tilde{\lambda}^T \tilde{R} \tilde{I}^{-1} \tilde{R}^T \tilde{\lambda} = \tilde{\sigma}^2 \tilde{\lambda}^T Q(X^T X)^{-1} Q^T \tilde{\lambda} = \\ &= \frac{1}{\tilde{\sigma}^2} (Q\hat{\beta} - q)^T (Q(X^T X)^{-1} Q^T)^{-1} (Q\hat{\beta} - q). \end{aligned}$$

Можно также показать (пропускаем эти преобразования), что

$$(Q\hat{\beta} - q)^T (Q(X^T X)^{-1} Q^T)^{-1} (Q\hat{\beta} - q) = R\tilde{S}S - R\hat{S}S.$$

Это позволяет выразить LM и W через суммы квадратов остатков:

$$LM = N \frac{R\tilde{S}S - R\hat{S}S}{R\tilde{S}S}, \quad W = N \frac{R\tilde{S}S - R\hat{S}S}{R\hat{S}S},$$

Логарифмическая функция правдоподобия в максимуме равна

$$\tilde{\ell} = -\frac{N}{2} \ln(2\pi \frac{R\tilde{S}S}{N}) - \frac{N}{2}.$$

В регрессии без ограничений  $\hat{\ell} = -\frac{N}{2} \ln(2\pi \frac{R\hat{S}S}{N}) - \frac{N}{2}$ .

Отсюда найдем статистику отношения правдоподобия:

$$LR = 2(\tilde{\ell} - \hat{\ell}) = N(\ln(R\tilde{S}S) - \ln(R\hat{S}S)).$$

Так как логарифм — строго вогнутая функция, то выполнено следующее точное неравенство:

$$W > LR > LM.$$

Таким образом, тест Вальда будет чаще отвергать гипотезу, тест множителя Лагранжа — реже. Это же неравенство верно и для нелинейных регрессий.

F-статистика для проверки той же гипотезы равна

$$F = \frac{R\tilde{SS} - R\hat{SS}}{R\hat{SS}} \frac{N-m}{p}.$$

Она распределена как  $F(p, N-m)$ .

В линейной регрессии лучше, конечно использовать t- и F-статистики. Кроме того, распределение этих статистик лучше аппроксимируется их номинальным распределением и в других моделях: нелинейных регрессиях, некоторых моделях, являющихся развитием регрессионных, некоторых искусственных регрессиях и т. п. Хотя здесь t- и F-статистики не будут иметь точного распределения, но, как показали эксперименты, они, как правило, лучше в конечных выборках, чем их асимптотические аналоги ( $N$  и  $\chi^2$  соответственно). Такие t- и F-статистики **называют асимптотическими t- и F-статистиками**. Три классические статистики можно преобразовать в асимптотические F-статистики по следующим формулам:

$$F_W = \frac{W}{N} \frac{N-m}{p}, \quad F_{LM} = \frac{LM}{N-LM} \frac{N-m}{p}, \quad F_{LR} = \left(\exp\left(\frac{LR}{N}\right) - 1\right) \frac{N-m}{p}.$$

Все эти статистики распределены приближенно как  $F(p, N-m)$ .

Понятно, что тесты на основе W, LR, LM и асимптотические F-тесты дают противоречивые результаты в конечных выборках. Одни из них могут отвергать гипотезу при выбранном уровне значимости, другие же говорить в пользу принятия гипотезы.

Для того, чтобы исследовать поведение асимптотических статистик в конечных выборках, используют метод Монте-Карло. С помощью этого метода можно, в частности, выяснить, какой из тестов более подходит для данного типа моделей, какую оценку ковариационной матрицы оценок лучше всего использовать.

## Модели с дискретной зависимой переменной

### Модели с бинарной зависимой переменной (логит и пробит)

Бинарная зависимая переменная  $Y_i$  называется так, потому, что принимает два значения, обычно 0 и 1. Обозначим через  $P_i$  вероятность появления единицы, или, что в данном случае то же самое, математическое ожидание  $Y_i$ :

$$P_i = \text{Prob}(Y_i = 1) = E(Y_i).$$

Вероятность  $P_i$  в линейной модели с бинарной зависимой переменной зависит от  $X_i \beta$ , где  $X_i$  — строка матрицы регрессоров,  $\beta$  — вектор коэффициентов регрессии:

$$P_i = F(X_i \beta).$$

Здесь  $F(\cdot)$  — (кумулятивная) функция распределения некоторого непрерывного распределения.

В **логите** используется (стандартное) логистическое распределение с функцией распределения

$$F(z) = \frac{1}{1 + e^{-z}}$$

и плотностью распределения

$$f(z) = \frac{e^z}{(1 + e^z)^2}.$$

В **пробите** используется стандартное нормальное распределение с функцией распределения

$$F(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt.$$

Логарифмическая функция правдоподобия равна

$$\begin{aligned} \ell &= \sum_{i \in I_1} \ln P_i(\beta) + \sum_{i \in I_0} \ln (1 - P_i(\beta)) = \\ &= \sum_{i=1}^N [Y_i \ln P_i(\beta) + (1 - Y_i) \ln (1 - P_i(\beta))]. \end{aligned}$$

где  $I_0$  и  $I_1$  — множества наблюдений, для которых  $Y_i = 0$  и  $Y_i = 1$  соответственно.

Градиент функции правдоподобия равен:

$$\mathbf{g}^\top = \frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_i \left[ \frac{Y_i}{P_i} - \frac{1-Y_i}{1-P_i} \right] \frac{\partial P_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_i \frac{Y_i - P_i}{P_i(1-P_i)} f(\mathbf{X}_i \boldsymbol{\beta}) \mathbf{X}_i.$$

### Логит

Для логита верно, что  $f(z) = F(z)(1-F(z))$ , поэтому  $f(\mathbf{X}_i \boldsymbol{\beta}) = P_i(1-P_i)$ . Это позволяет упростить формулу градиента:

$$\mathbf{g}^\top = \sum_i (Y_i - P_i) \mathbf{X}_i,$$

где  $P_i = \frac{1}{1 + e^{-\mathbf{X}_i \boldsymbol{\beta}}}$ .

Гессиан в случае логита равен:

$$\begin{aligned} \mathcal{H} &= \frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = - \sum_i \frac{\partial P_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top} \mathbf{X}_i^\top = - \sum_i f(\mathbf{X}_i \boldsymbol{\beta}) \mathbf{X}_i^\top \mathbf{X}_i = \\ &= - \sum_i P_i(1-P_i) \mathbf{X}_i^\top \mathbf{X}_i. \end{aligned}$$

Видно, что гессиан всюду отрицательно определенный (кроме вырожденных случаев). Таким образом, логарифмическая функция правдоподобия всюду вогнута.

Гессиан не зависит от случайного вектора  $\mathbf{Y}$ , поэтому ожидаемый гессиан равен просто гессиану, то есть информационная матрица равна минус гессиану:

$$\mathcal{I} = -\mathcal{H} = \sum_i P_i(1-P_i) \mathbf{X}_i^\top \mathbf{X}_i.$$

Для поиска максимума можно использовать метод Ньютона (он же в данном случае и method of scoring):

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - (\mathcal{H}(\boldsymbol{\beta}^t))^{-1} \mathbf{g}^t(\boldsymbol{\beta}^t) = \boldsymbol{\beta}^t - \Delta \boldsymbol{\beta}^t.$$

Поскольку максимизируемая функция вогнута, то метод Ньютона всегда сходится. Шаг алгоритма удобно находить как оценки коэффициентов во вспомогательной регрессии  $\mathbf{Y}^*$  по  $\mathbf{X}^*$ , где

$$\mathbf{Y}_i^* = \frac{Y_i - P_i}{\sqrt{P_i(1-P_i)}}, \quad \mathbf{X}_i^* = \sqrt{P_i(1-P_i)} \mathbf{X}_i.$$

### Пробит

В случае пробита выражение для гессиана несколько более громоздкое:

$$\begin{aligned} \mathcal{H} &= \frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = \frac{\partial}{\partial \boldsymbol{\beta}^\top} \sum_i \left[ \frac{Y_i}{P_i} - \frac{1-Y_i}{1-P_i} \right] f(\mathbf{X}_i \boldsymbol{\beta}) \mathbf{X}_i = \\ &= - \sum_i \left[ \frac{Y_i}{(P_i)^2} + \frac{1-Y_i}{(1-P_i)^2} \right] f(\mathbf{X}_i \boldsymbol{\beta})^2 \mathbf{X}_i^\top \mathbf{X}_i \end{aligned}$$

$$\begin{aligned}
& + \sum_i \left[ \frac{Y_i}{P_i} - \frac{1-Y_i}{1-P_i} \right] f(\mathbf{X}_i \boldsymbol{\beta}) \frac{\partial f(\mathbf{X}_i \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top} \mathbf{X}_i = \\
& = - \sum_i \left[ \left( \frac{Y_i - P_i}{P_i(1-P_i)} \right)^2 f(\mathbf{X}_i \boldsymbol{\beta}) \mathbf{X}_i^\top - \frac{Y_i - P_i}{P_i(1-P_i)} \frac{\partial f(\mathbf{X}_i \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top} \right] f(\mathbf{X}_i \boldsymbol{\beta}) \mathbf{X}_i.
\end{aligned}$$

Для нормального распределения верно, что  $\frac{\partial f(z)}{\partial z} = -zf(z)$ . Это позволяет несколько упростить выражение для гессиана, так как

$$\frac{\partial f(\mathbf{X}_i \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top} = -\mathbf{X}_i \boldsymbol{\beta} f(\mathbf{X}_i \boldsymbol{\beta}) \mathbf{X}_i^\top.$$

Обозначим

$$v_i = f(\mathbf{X}_i \boldsymbol{\beta}) \frac{Y_i - P_i}{P_i(1-P_i)}.$$

Тогда

$$\mathcal{H} = - \sum_i v_i (v_i + \mathbf{X}_i \boldsymbol{\beta}) \mathbf{X}_i^\top \mathbf{X}_i.$$

В тех же обозначениях градиент равен

$$\mathbf{g}^\top = \sum_i v_i \mathbf{X}_i.$$

Как и в случае логита, можно показать, что гессиан является отрицательно определенным.

Чтобы найти информационную матрицу для пробита, воспользуемся тем, что  $E(Y_i) = P_i$ ,  $E(Y_i - P_i)^2 = P_i(1 - P_i)$ .

$$\mathcal{I} = -E(\mathcal{H}) = - \sum_i E(v_i^2) \mathbf{X}_i^\top \mathbf{X}_i = - \sum_i \frac{f^2(\mathbf{X}_i \boldsymbol{\beta})}{P_i(1-P_i)} \mathbf{X}_i^\top \mathbf{X}_i.$$

Для поиска максимума, как и в случае логита, можно использовать градиентный алгоритм:

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - (\mathcal{I}^t)^{-1} \mathbf{g}^t = \boldsymbol{\beta}^t - \Delta \boldsymbol{\beta}^t.$$

В методе Ньютона с  $\mathcal{I}^t = -\mathcal{H}(\boldsymbol{\beta}^t)$  используется вспомогательная регрессия с переменными

$$Y_i^* = \frac{\sqrt{v_i}}{\sqrt{v_i + \mathbf{X}_i \boldsymbol{\beta}}}, \quad X_i^* = \sqrt{v_i(v_i + \mathbf{X}_i \boldsymbol{\beta})} \mathbf{X}_i.$$

Если использовать информационную матрицу в точке оценок  $\mathcal{I}^t = \mathcal{I}(\boldsymbol{\beta}^t)$  (method of scoring), то надо взять

$$Y_i^* = \frac{Y_i - P_i}{\sqrt{P_i(1-P_i)}}, \quad X_i^* = \frac{f(\mathbf{X}_i \boldsymbol{\beta})}{\sqrt{P_i(1-P_i)}} \mathbf{X}_i.$$

Вспомогательные регрессии для пробита и логита являются искусственными регрессиями, то есть, с помощью них можно проверять все те гипотезы, которые можно проверять в случае обычной регрессии, в частности, использовать t-статистики.

Метод максимального правдоподобия для моделей с дискретной зависимой переменной по сути является нелинейным методом наименьших квадратов (НМНК). Математическое ожидание  $Y_i$  равно  $P_i$ . Разность  $Y_i$  и  $P_i$  должна иметь нулевое математическое ожидание, то есть подходит в качестве ошибки в нелинейной регрессии  $Y_i$  по  $P_i$ . Однако эта ошибка будет гетероскедастична. Действительно,

$$V(Y_i) = E(Y_i - P_i)^2 = P_i(1 - P_i)^2 + (1 - P_i)P_i^2 = P_i(1 - P_i).$$

Таким образом, следует воспользоваться взвешенным НМНК, где веса рассматриваются как фиксированные:

$$\sum_{i=1}^N \frac{(Y_i - P_i)^2}{P_i(1 - P_i)} \rightarrow \min.$$

Поскольку веса неизвестны, то приходится использовать итерационные процедуры, которые совпадают с описанными выше. Оба метода дают одни и те же оценки, поскольку целевые функции достигают экстремума одновременно.

### Пуассоновская регрессия

Распределение Пуассона — дискретное распределение, задаваемое формулой

$$\text{Prob}(Y=r) = \frac{\mu^r}{r!} e^{-\mu},$$

где  $\mu$  — параметр распределения.

Распределение Пуассона имеет случайная величина  $Y$ , равная количеству событий, произошедших за некоторый промежуток времени, если эти события независимы и происходят с постоянной скоростью (равномерно по времени). Это, например, может быть количество покупателей, посетивших магазин в течении часа.

Моменты распределения:

$$E(Y) = \mu, \quad \text{Var}(Y) = \mu.$$

В регрессионной модели с распределением Пуассона параметр  $\mu$  зависит от набора факторов и неизвестных параметров.

В линейной модели:

$$\mu_i = \exp(\mathbf{X}_i \boldsymbol{\beta}).$$

Тогда логарифмическая функция правдоподобия равна

$$\ell = \sum_i [Y_i \mathbf{X}_i \boldsymbol{\beta} - \exp(\mathbf{X}_i \boldsymbol{\beta}) - \ln Y_i!] \rightarrow \max_{\boldsymbol{\beta}}.$$

Градиент равен:

$$\mathbf{g}^T = \frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_i [Y_i \mathbf{X}_i - \exp(\mathbf{X}_i \boldsymbol{\beta}) \mathbf{X}_i].$$

Условие первого порядка максимума:

$$\sum_i [Y_i - \exp(\mathbf{X}_i \boldsymbol{\beta})] \mathbf{X}_i = 0.$$

Гессиан не содержит случайных компонент, и поэтому информационная матрица равна минус гессиану.

$$\mathcal{H} = \frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = - \sum_i \exp(\mathbf{X}_i \boldsymbol{\beta}) \mathbf{X}_i^T \mathbf{X}_i.$$

Для поиска максимума можно использовать метод Ньютона:

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - (\mathcal{H}^t)^{-1} \mathbf{g}^t.$$

Метод Ньютона легко реализовать с помощью вспомогательной регрессии.

Обозначим

$$v_i = \exp\left(\frac{1}{2} \mathbf{X}_i \boldsymbol{\beta}\right), \quad Y_i^* = Y_i / v_i - v_i, \quad X_i^* = X_i v_i.$$

Тогда если  $\Delta \boldsymbol{\beta}$  — оценки коэффициентов в регрессии  $Y^*$  по  $X^*$ , то шаг метода Ньютона задается формулой:

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - \Delta \boldsymbol{\beta}^t.$$

Оценка ковариационной матрицы оценок есть  $-(\mathcal{H})^{-1} = (\mathbf{X}^{*T} \mathbf{X}^*)^{-1}$ , поэтому тесты для коэффициентов матрицы и т. п. можно получить из регрессии  $Y^{**}$  по  $X^*$ , где  $Y^{**} = Y^*/s$ ,  $s = \sqrt{Y^{*T} Y^*/N}$ , и  $Y^*$  берется в точке оценок МП (на последней итерации метода Ньютона). Проверять ограничения на коэффициенты можно как с помощью  $\chi^2$  статистики Вальда, так и с помощью соответствующих t- и F-статистик из вспомогательной регрессии. В качестве аналога стандартного F-теста на равенство нулю коэффициентов при всех переменных кроме константы можно использовать статистику отношения правдоподобия. Пусть  $\tilde{\ell}$  — значение логарифмической функции правдоподобия когда  $\mu_i = \mu \forall i$ .

$$\text{LR} = 2 (\hat{\ell} - \tilde{\ell})^a \sim \chi^2(m-1),$$

где  $m$  — количество регрессоров (столбцов  $\mathbf{X}$ ).

Найдем  $\tilde{\ell}$ :

$$\ell = \sum_i [Y_i \ln \mu - \mu - \ln Y_i!].$$

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\mu} \sum_i Y_i - N = 0 \quad \Rightarrow \quad \tilde{\mu} = \frac{1}{N} \sum_i Y_i = \bar{Y}.$$

Таким образом,  $\tilde{\ell} = N \bar{Y} \ln \bar{Y} - N \bar{Y} - \sum_i \ln Y_i!$ .

## Обобщенный метод наименьших квадратов

Отбросим предположение, что в регрессионной модели ошибки независимы. Пусть ошибки коррелированы и структура матрицы ковариаций ошибок  $V$  известна. Найдем оценки в этой регрессии методом МП при нормально распределенных ошибках. Заметим, что метод можно использовать и для более широкого класса распределений.

Модель имеет вид

$$Y = X\beta + \epsilon, \text{ где } \epsilon \sim N(0, V).$$

Плотность распределения  $\epsilon$  равна

$$p_{\epsilon}(z) = (2\pi)^{-N/2} |V|^{-1/2} \exp\left(-\frac{1}{2} z^T V z\right).$$

Отсюда получим логарифмическую функцию правдоподобия

$$\ell = -\frac{N}{2} \ln(2\pi) - \frac{1}{2} \ln|V| - \frac{1}{2} (Y - X\beta)^T V^{-1} (Y - X\beta).$$

Далее мы рассмотрим случай, когда  $V$  известна с точностью до множителя:

$$V = \sigma^2 W.$$

Подставим это выражение в функцию правдоподобия:

$$\ell = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \ln|W| - \frac{1}{2\sigma^2} (Y - X\beta)^T W^{-1} (Y - X\beta).$$

Воспользовавшись условием первого порядка

$$\frac{\partial \ell}{\partial \sigma^2} = 0,$$

выразим  $\sigma^2$  через  $\beta$ :  $\sigma^2(\beta) = \text{RSS}(\beta)/N$  где

$$\text{RSS}(\beta) = (Y - X\beta)^T W^{-1} (Y - X\beta)$$

— обобщенная сумма квадратов остатков.

Подставив  $\sigma^2(\beta)$  в логарифмическую функцию правдоподобия, получим концентрированную функцию правдоподобия:

$$\ell^c = -\frac{N}{2} \ln\left(2\pi \frac{\text{RSS}(\beta)}{N}\right) - \frac{1}{2} \ln|W| - \frac{N}{2}.$$

Поскольку  $W$  — константа, то максимизация  $\ell^c$  эквивалентна минимизации обобщенной суммы квадратов:  $\text{RSS}(\beta) \rightarrow \min_{\beta}$ .

$$\frac{\partial \ell^c}{\partial \beta} = 0 \Leftrightarrow Y^T W^{-1} X = \hat{\beta}^T X^T W^{-1} X.$$

Получим оценку МП для  $\beta$ :

$$\hat{\beta} = (X^T W^{-1} X)^{-1} X^T W^{-1} Y.$$

На практике удобно использовать не эту формулу, а преобразовать  $X$  и  $Y$  так, чтобы можно было делать расчеты с помощью ОМНК. Поскольку  $W$  — симметричная положительно определенная матрица, то к ней можно применить разложение Холецкого (или любое другое аналогичное представление):

$$W = T T^T,$$

где  $T$  — нижняя или верхняя треугольная матрица. Отсюда

$$W^{-1} = (T^{-1})^T T^{-1},$$

$$\hat{\beta} = (X^T (T^{-1})^T T^{-1} X)^{-1} X^T (T^{-1})^T T^{-1} Y.$$

Обозначим  $T^{-1} X = X^*$ ,  $T^{-1} Y = Y^*$ . Тогда выражение для  $\hat{\beta}$  примет вид:

$$\hat{\beta} = (X^{*T} X^*)^{-1} X^{*T} Y^*.$$

Таким образом, оценки коэффициентов  $\beta$  можно найти, применив обычный метод наименьших квадратов к регрессии  $Y^*$  по  $X^*$ . Вообще говоря, предположения о нормальности не требуется для состоятельности оценок, и на метод максимального правдоподобия можно было не ссылаться, поскольку предложенное преобразование сразу приводит к классической регрессионной модели.

Несложно проверить, что информационная матрица (для  $\beta$ ) равна

$$I = \frac{1}{\sigma^2} X^T W^{-1} X,$$

поэтому и ковариационная матрица оценок  $\hat{\beta}$  полученная из той же регрессии будет оценкой ковариационной матрицы оценок максимального правдоподобия:

$$V(\hat{\beta}) = \hat{\sigma}^2 (X^T W^{-1} X)^{-1} = \hat{\sigma}^2 (X^{*T} X^*)^{-1},$$

где  $\hat{\sigma}^2$  — оценка дисперсии в регрессии  $Y^*$  по  $X^*$ , которая является оценкой параметра  $\sigma^2$  в исходной модели. Это позволяет использовать t- и F-статистики в преобразованной модели для проверки гипотез о коэффициентах  $\beta$ .

В более общем случае матрица ковариаций ошибок  $V$  зависит от вектора неизвестных параметров  $\alpha$ . Эту модель в дальнейшем будем называть моделью обобщенного метода наименьших квадратов, хотя ковариационная матрица ошибок в обобщенном методе наименьших квадратов в собственном

смысле слова зависит от единственного неизвестного множителя. Предполагается, что два вектора параметров не связаны между собой, т.е. математическое ожидание  $Y$  не зависит от  $\alpha$ , а матрица ковариаций  $V$  не зависит от  $\beta$ .

$$\ell = -\frac{N}{2} \ln(2\pi) - \frac{1}{2} \ln |V(\alpha)| - \frac{1}{2} (Y - X\beta)^\top V(\alpha)^{-1} (Y - X\beta).$$

Информационная матрица будет блочно-диагональной: ее часть, соответствующая “взаимодействию”  $\alpha$  и  $\beta$  будет равна нулю:

$$\frac{\partial \ell}{\partial \beta} = (Y - X\beta)^\top V(\alpha)^{-1} X.$$

$$\frac{\partial^2 \ell}{\partial \beta \partial \alpha^\top} = (Y - X\beta)^\top \frac{\partial V(\alpha)^{-1}}{\partial \alpha^\top} X.$$

$$\mathcal{I}_{\beta\alpha} = E\left(\frac{\partial^2 \ell}{\partial \beta \partial \alpha^\top}(\beta_0, \alpha_0)\right) = E(\epsilon^\top \frac{\partial V(\alpha)^{-1}}{\partial \alpha^\top} X) = E(\epsilon^\top) \frac{\partial V(\alpha)^{-1}}{\partial \alpha^\top} X = \mathbf{O}.$$

Отсюда следует, что для проведения тестов относительно  $\alpha$  можно использовать просто диагональный блок информационной матрицы, относящийся к  $\alpha$ , не учитывая, что  $\hat{\beta}$  — оценки, и наоборот, для проведения тестов относительно  $\beta$  можно использовать просто диагональный блок информационной матрицы, относящийся к  $\beta$ :

$$\mathcal{I}_{\beta\beta} = E\left(\frac{\partial \ell}{\partial \beta^\top} \frac{\partial \ell}{\partial \beta}\right) = X^\top V(\alpha)^{-1} X.$$

Если имеется способ получить состоятельную оценку параметров  $\alpha$  ( $\bar{\alpha}$ ), то эффективную оценку параметров  $\beta$  благодаря блочно-диагональности информационной матрицы можно получить за один шаг:

$$\bar{\beta} = (X^\top V(\bar{\alpha})^{-1} X)^{-1} X^\top V(\bar{\alpha})^{-1} Y.$$

Такую оценку принято называть **одношаговой эффективной оценкой**, а метод называют **возможный обобщенный метод наименьших квадратов** (feasible generalized least squares).

Если на основании оценок  $\hat{\beta}$  можно из условий первого порядка максимума правдоподобия вычислить оценки  $\alpha$ , то можно использовать **итеративный обобщенный метод наименьших квадратов** (iterated generalized least squares). Этот метод сходится к оценкам максимального правдоподобия.

## Регрессии с неодинаковой дисперсией и тестирование гетероскедастичности

### Взвешенный метод наименьших квадратов

Обобщенный метод наименьших квадратов имеет много применений. Его частным случаем является взвешенный метод наименьших квадратов, позволяющий оценивать регрессии с гетероскедастичной ошибкой. Гетероскедастичность означает, что хотя матрица ковариаций ошибок диагональная, но дисперсии (стоящие по диагонали) разные.

Пусть ошибки независимы и  $i$ -я ошибка имеет дисперсию  $\sigma_i^2 = \sigma^2 w_i$ . В данном случае матрица  $\mathbf{W}$  — диагональная с типичным диагональным элементом  $w_i$ . Матрица  $\mathbf{T}$  — тоже диагональная с типичным элементом  $\sqrt{w_i}$ , а  $\mathbf{T}^{-1}$  — диагональная с типичным элементом  $\frac{1}{\sqrt{w_i}}$ . Переменные во вспомогательной регрессии будут иметь вид:

$$Y_i^* = \frac{Y_i}{\sqrt{w_i}}, \quad X_i^* = \frac{X_i}{\sqrt{w_i}}.$$

Такую регрессию называют **взвешенной регрессией**.

Если веса зависят от неизвестных параметров  $w_i = w_i(\boldsymbol{\gamma})$ , то следует воспользоваться методом максимального правдоподобия. Логарифмическая функция правдоподобия равна

$$\ell = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \sum_i \ln w_i(\boldsymbol{\gamma}) - \frac{1}{2\sigma^2} \sum_i \frac{1}{w_i(\boldsymbol{\gamma})} (Y_i - \mathbf{X}_i \boldsymbol{\beta})^2.$$

Концентрируем функцию правдоподобия по  $\sigma^2$ :

$$\ell^c = -\frac{1}{2} \sum_i \ln \left( \frac{1}{w_i(\boldsymbol{\gamma})} (Y_i - \mathbf{X}_i \boldsymbol{\beta})^2 \right) - \frac{1}{2} \sum_i \ln w_i(\boldsymbol{\gamma}) + \text{const}.$$

Максимизация функции правдоподобия эквивалентна минимизации суммы квадратов остатков взвешенной регрессии по  $\boldsymbol{\beta}$  и  $\boldsymbol{\gamma}$ , если взять нормированные веса:

$$Y_i^n = \frac{Y_i}{\sqrt{w_i^n(\boldsymbol{\gamma})}}, \quad X_i^n = \frac{X_i}{\sqrt{w_i^n(\boldsymbol{\gamma})}}.$$

Здесь  $w_i^n(\boldsymbol{\gamma}) = \frac{w_i(\boldsymbol{\gamma})}{\bar{w}(\boldsymbol{\gamma})}$ ,  $\bar{w}(\boldsymbol{\gamma})$  — среднее геометрическое весов ( $\bar{w} = \prod_i w_i^{1/N}$ ).

Важно, что используются *нормированные* веса, в противном случае минимизация суммы квадратов привела бы к неправильному результату.

Такой метод малоприменим для вычислений. Ниже рассматривается более удобный метод, который годится в частном случае линейной мультипликативной гетероскедастичности.

### Проверка гипотезы о наличии гетероскедастичности известного вида

Выдвинем явную гипотезу о виде гетероскедастичности в регрессии:

$$w_i(\boldsymbol{\gamma}) = h(\mathbf{Z}_i\boldsymbol{\gamma}),$$

где  $h(\cdot)$  — дифференцируемая строго монотонная функция, такая что  $h(0) = 1$ ,  $\mathbf{Z}_i\boldsymbol{\gamma}$  — линейная комбинация известных переменных  $\mathbf{Z}$  с неизвестными коэффициентами  $\boldsymbol{\gamma}$ .

Дисперсия ошибки  $i$ -го наблюдения равна  $\sigma_i^2 = \sigma^2 h(\mathbf{Z}_i\boldsymbol{\gamma})$ . Функция правдоподобия  $i$ -го наблюдения будет иметь вид:

$$\ell_i = -\frac{1}{2} \ln(2\pi\sigma^2 h(\mathbf{Z}_i\boldsymbol{\gamma})) - \frac{1}{2\sigma^2 h(\mathbf{Z}_i\boldsymbol{\gamma})} (Y_i - \mathbf{X}_i\boldsymbol{\beta})^2.$$

Как мы уже видели, информационная матрица в модели обобщенного МНК имеет блочно-диагональную форму, поэтому гипотезы о  $\boldsymbol{\gamma}$  можно проверять независимо от  $\boldsymbol{\beta}$ . Поэтому в дальнейшем будем рассматривать градиент функции правдоподобия и информационную матрицу только в той части, которая относится к  $\boldsymbol{\gamma}$  и  $\sigma^2$ , которые вместе составляют вектор  $\boldsymbol{\alpha} = (\sigma^2, \boldsymbol{\gamma})^T$ .

Для проверки гипотезы об отсутствии гетероскедастичности удобнее всего использовать LM-тест (нулевая гипотеза  $H_0: \boldsymbol{\gamma} = \mathbf{0}$ ), поскольку для него не требуется оценивать модель при  $\boldsymbol{\gamma} \neq \mathbf{0}$ . Достаточно оценить регрессию обычным методом наименьших квадратов.

Найдем вклад в градиент  $i$ -го наблюдения:

$$\frac{\partial \ell_i}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4} \frac{\varepsilon_i^2}{h(\mathbf{Z}_i\boldsymbol{\gamma})}.$$

$$\left. \frac{\partial \ell_i}{\partial \sigma^2} \right|_{H_0} = \frac{1}{2\sigma^2} \left( \frac{\varepsilon_i^2}{\sigma^2} - 1 \right) = \frac{1}{2\sigma^2} \mu_i.$$

$$\frac{\partial \ell_i}{\partial \boldsymbol{\gamma}} = -\frac{1}{2} \frac{h'(\mathbf{Z}_i\boldsymbol{\gamma})}{h(\mathbf{Z}_i\boldsymbol{\gamma})} \mathbf{Z}_i + \frac{1}{2\sigma^2} \frac{h'(\mathbf{Z}_i\boldsymbol{\gamma})}{h^2(\mathbf{Z}_i\boldsymbol{\gamma})} \varepsilon_i^2 \mathbf{Z}_i.$$

$$\left. \frac{\partial \ell_i}{\partial \boldsymbol{\gamma}} \right|_{H_0} = \frac{h'(0)}{2} \left( \frac{\varepsilon_i^2}{\sigma^2} - 1 \right) \mathbf{Z}_i = \frac{h'(0)}{2} \mu_i \mathbf{Z}_i.$$

Здесь мы обозначили  $\mu_i = \frac{\varepsilon_i^2}{\sigma^2} - 1$  и воспользовались тем, что  $h(0) = 1$ . Информационную матрицу удобно находить через матрицу вкладов в градиент. Воспользуемся тем, что  $E(\mu_i^2) = 2$ , поскольку для нормального распределения

$$E\left(\frac{\varepsilon_i^2}{\sigma^2}\right) = 1 \quad \text{и} \quad E\left(\frac{\varepsilon_i^4}{\sigma^4}\right) = 3.$$

Отсюда получим при выполнении нулевой гипотезы

$$\begin{aligned} E\left(\left(\frac{\partial \ell_i}{\partial \sigma^2}\right)^2\right) &= \frac{1}{4\sigma_i^4} E(\mu_i^2) = \frac{1}{2\sigma_i^4}, \\ E\left(\frac{\partial \ell_i}{\partial \sigma^2} \frac{\partial \ell_i}{\partial \boldsymbol{\gamma}}\right) &= \frac{h'(0)}{4\sigma^2} E(\mu_i^2) \mathbf{Z}_i = \frac{h'(0)}{2\sigma^2} \mathbf{Z}_i, \\ E\left(\frac{\partial \ell_i}{\partial \boldsymbol{\gamma}^\top} \frac{\partial \ell_i}{\partial \boldsymbol{\gamma}}\right) &= \frac{(h'(0))^2}{4} E(\mu_i^2) \mathbf{Z}_i^\top \mathbf{Z}_i = \frac{(h'(0))^2}{2} \mathbf{Z}_i^\top \mathbf{Z}_i. \end{aligned}$$

Таким образом, информационная матрица равна

$$\begin{aligned} \mathcal{I}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^N &= E(\mathbf{G}_{\boldsymbol{\alpha}0}^\top \mathbf{G}_{\boldsymbol{\alpha}0}) = \begin{bmatrix} \frac{N}{2\sigma_i^4} & \frac{h'(0)}{2\sigma^2} \sum_i \mathbf{Z}_i \\ \frac{h'(0)}{2\sigma^2} \sum_i \mathbf{Z}_i^\top & \frac{(h'(0))^2}{2} \sum_i \mathbf{Z}_i^\top \mathbf{Z}_i \end{bmatrix} = \\ &= \begin{bmatrix} \frac{1}{2\sigma_i^4} \mathbf{1}^\top \mathbf{1} & \frac{h'(0)}{2\sigma^2} \mathbf{1}^\top \mathbf{Z} \\ \frac{h'(0)}{2\sigma^2} \mathbf{Z}^\top \mathbf{1} & \frac{(h'(0))^2}{2} \mathbf{Z}^\top \mathbf{Z} \end{bmatrix}. \end{aligned}$$

где  $\mathbf{1}$  — вектор-столбец, составленный из  $N$  единиц. Если обозначить

$$\mathbf{Z}^* = \left( \frac{1}{\sqrt{2}\sigma^2} \mathbf{1}, \frac{h'(0)}{\sqrt{2}} \mathbf{Z} \right),$$

то

$$\mathcal{I}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^N = \mathbf{Z}^{*\top} \mathbf{Z}^*.$$

Статистика множителя Лагранжа для проверяемой гипотезы равна

$$\text{LM} = \tilde{\mathbf{g}}_{\boldsymbol{\alpha}}^\top (\tilde{\mathcal{I}}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^N)^{-1} \tilde{\mathbf{g}}_{\boldsymbol{\alpha}},$$

где градиент и информационная матрица берутся в точке  $(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, 0)$  оценок ОМНК.

Градиент равен  $\tilde{\mathbf{g}}_{\alpha} = \left( \frac{1}{2\tilde{\sigma}^2} \mathbf{1}^T \tilde{\boldsymbol{\mu}}, \frac{h'(0)}{2} \mathbf{Z}^T \tilde{\boldsymbol{\mu}} \right)$ , где  $\tilde{\mu}_i = \frac{e_i^2}{\tilde{\sigma}^2} - 1$ ,  $e_i$  — остатки из регрессии. (Оценка дисперсии  $\tilde{\sigma}^2$ , полученная методом максимального правдоподобия такова, что  $\mathbf{1}^T \tilde{\boldsymbol{\mu}} = 0$ , так как производная функции правдоподобия равна нулю.) В терминах матрицы  $\mathbf{Z}^*$

$$\tilde{\mathbf{g}}_{\alpha} = \frac{1}{\sqrt{2}} \mathbf{Z}^{*T} \tilde{\boldsymbol{\mu}}.$$

В таком случае можно заметить, что LM-статистика равна *объясненной сумме квадратов* из регрессии  $\frac{1}{\sqrt{2}} \tilde{\boldsymbol{\mu}}$  по  $\mathbf{Z}^*$  или, что то же самое, *половине объясненной суммы квадратов* из регрессии  $\tilde{\boldsymbol{\mu}}$  по  $\mathbf{Z}^*$ :

$$\text{LM} = \frac{1}{\sqrt{2}} \tilde{\boldsymbol{\mu}}^T \mathbf{Z}^* (\mathbf{Z}^{*T} \mathbf{Z}^*)^{-1} \frac{1}{\sqrt{2}} \mathbf{Z}^{*T} \tilde{\boldsymbol{\mu}} = \frac{1}{2} \tilde{\boldsymbol{\mu}}^T \mathbf{Z}^* (\mathbf{Z}^{*T} \mathbf{Z}^*)^{-1} \mathbf{Z}^{*T} \tilde{\boldsymbol{\mu}}.$$

Если домножить регрессоры на отличные от нуля константы, то подпространство, которое на них натянуто, не изменится. Поэтому регрессия  $\tilde{\boldsymbol{\mu}}$  по  $\mathbf{Z}^*$  дает ту же самую объясненную сумму квадратов, что и регрессия  $\tilde{\boldsymbol{\mu}}$  по  $\mathbf{1}$  и  $\mathbf{Z}$ . Таким образом, окончательно получаем, что LM-статистика для тестирования гетероскедастичности равна половине объясненной суммы квадратов из регрессии  $\tilde{\boldsymbol{\mu}}$  по константе и  $\mathbf{Z}$ . Статистика распределена асимптотически как  $\chi^2(r)$ , где  $r$  — размерность вектора  $\boldsymbol{\gamma}$ .

Примечательно, что в этой статистике не фигурируют производные функции  $h(\cdot)$ , формула будет одна и та же независимо от выбора  $h(\cdot)$ . Когда статистика множителя Лагранжа одна и та же для широкого класса альтернативных гипотез, тогда эти альтернативные модели принято называть **локально эквивалентными альтернативами**.

### Регрессия с мультипликативной гетероскедастичностью

В регрессии с (линейной) **мультипликативной гетероскедастичностью** дисперсия ошибки равна

$$\sigma_i^2(\boldsymbol{\alpha}) = \exp(\mathbf{Z}_i \boldsymbol{\alpha}).$$

Здесь  $\mathbf{Z}$  — матрица, состоящая из переменных, от которых зависит дисперсия (как правило, в ней должен быть столбец, состоящий из единиц),  $\boldsymbol{\alpha}$  — вектор параметров.

Регрессия задана формулой:

$$Y_i = \mathbf{X}_i \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_i \sim \text{NID}(0, \sigma_i^2(\boldsymbol{\alpha})).$$

Предполагается, что неизвестные параметры в “среднем” и в дисперсии не связаны между собой.

Логарифмическая функция правдоподобия  $i$ -го наблюдения для этой модели имеет вид:

$$\begin{aligned} \ell_i &= -\frac{1}{2} \ln(2\pi\sigma_i^2(\boldsymbol{\alpha})) - \frac{1}{2\sigma_i^2(\boldsymbol{\alpha})} (Y_i - \mathbf{X}_i \boldsymbol{\beta})^2 = \\ &= -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \mathbf{Z}_i \boldsymbol{\alpha} - \frac{e_i^2}{2 \exp(\mathbf{Z}_i \boldsymbol{\alpha})}. \end{aligned}$$

Найдем вклад в градиент  $i$ -го наблюдения:

$$\begin{aligned} \frac{\partial \ell_i}{\partial \boldsymbol{\beta}} &= \frac{1}{\sigma_i^2(\boldsymbol{\alpha})} e_i \mathbf{X}_i, \\ \frac{\partial \ell_i}{\partial \boldsymbol{\alpha}} &= -\frac{1}{2} \mathbf{Z}_i + \frac{e_i^2}{2 \exp(\mathbf{Z}_i \boldsymbol{\alpha})} \mathbf{Z}_i = \frac{1}{2} \left( \frac{e_i^2}{\sigma_i^2(\boldsymbol{\alpha})} - 1 \right) \mathbf{Z}_i. \end{aligned}$$

Вклад в информационную матрицу  $i$ -го наблюдения равен

$$\begin{aligned} \mathbb{E} \left( \frac{\partial \ell_i}{\partial \boldsymbol{\beta}^\top} \frac{\partial \ell_i}{\partial \boldsymbol{\beta}} \right) &= \frac{\mathbb{E}(\varepsilon_i^2)}{\sigma_i^4} \mathbf{X}_i^\top \mathbf{X}_i = \frac{1}{\sigma_i^2} \mathbf{X}_i^\top \mathbf{X}_i, \\ \mathbb{E} \left( \frac{\partial \ell_i}{\partial \boldsymbol{\beta}^\top} \frac{\partial \ell_i}{\partial \boldsymbol{\alpha}} \right) &= \frac{1}{2} \mathbb{E} \left( \frac{\varepsilon_i}{\sigma_i^2(\boldsymbol{\alpha})} \left( \frac{\varepsilon_i^2}{\sigma_i^2(\boldsymbol{\alpha})} - 1 \right) \right) \mathbf{X}_i^\top \mathbf{Z}_i = 0, \\ \mathbb{E} \left( \frac{\partial \ell_i}{\partial \boldsymbol{\alpha}^\top} \frac{\partial \ell_i}{\partial \boldsymbol{\alpha}} \right) &= \frac{1}{4} \mathbb{E} \left( \frac{\varepsilon_i^4}{\sigma_i^4(\boldsymbol{\alpha})} - 2 \frac{\varepsilon_i^2}{\sigma_i^2(\boldsymbol{\alpha})} + 1 \right) \mathbf{Z}_i^\top \mathbf{Z}_i = \\ &= \frac{1}{4} (3 - 2 + 1) \mathbf{Z}_i^\top \mathbf{Z}_i = \frac{1}{2} \mathbf{Z}_i^\top \mathbf{Z}_i. \end{aligned}$$

Таким образом, информационная матрица (как и следовало ожидать) блочно-диагональная и блоки ее равны:

$$\mathcal{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}^N = \mathbf{X}^\top \text{diag} \left( \frac{1}{\sigma_1^2}, \dots, \frac{1}{\sigma_N^2} \right) \mathbf{X}, \quad \mathcal{I}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^N = \frac{1}{2} \mathbf{Z}^\top \mathbf{Z}.$$

При данном векторе  $\boldsymbol{\alpha}$ , коэффициенты регрессии  $\boldsymbol{\beta}$  можно найти из взвешенной регрессии:

$$\boldsymbol{\beta} = (\mathbf{X}^{*\top} \mathbf{X}^*)^{-1} \mathbf{X}^{*\top} \mathbf{Y}^*,$$

где  $X_i^* = \frac{1}{\sigma_i} X_i$ ,  $Y_i^* = \frac{1}{\sigma_i} Y_i$ . Обозначим остатки из этой регрессии  $e^*(\boldsymbol{\alpha})$ .

Используем итерации по  $\boldsymbol{\alpha}$ :

$$\boldsymbol{\alpha}^{t+1} = \boldsymbol{\alpha}^t + \mathcal{I}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^N(\boldsymbol{\alpha}^t)^{-1} \mathbf{g}^{\boldsymbol{\alpha}} = \boldsymbol{\alpha}^t + \Delta\boldsymbol{\alpha}^t.$$

$\Delta\boldsymbol{\alpha}^t$  можно находить с помощью вспомогательной регрессии  $\frac{1}{\sqrt{2}} \tilde{\boldsymbol{\mu}}(\boldsymbol{\alpha}^t)$  по  $\frac{1}{\sqrt{2}} \mathbf{Z}$ , где  $\tilde{\mu}_i = \frac{e_i^2}{\sigma_i^2} - 1 = (e_i^*)^2 - 1$ .

Обе используемые в этом алгоритме вспомогательные регрессии дают состоятельные оценки ковариационных матриц соответствующих оценок параметров и могут использоваться для проверки гипотез.

## **Нелинейная регрессия. Метод Гаусса-Ньютона**

Пусть нелинейная регрессия задана уравнением

$$Y_i = f_i(\boldsymbol{\theta}) + \varepsilon_i,$$

Имея на  $t$ -м шаге приближение  $\boldsymbol{\theta}^t$ , следующее приближение  $\boldsymbol{\theta}^{t+1}$  получаем с помощью регрессии  $Y - \mathbf{f}(\boldsymbol{\theta}^t)$  по  $\mathbf{F}(\boldsymbol{\theta}^t)$ , где  $\mathbf{F}(\cdot)$  — матрица производных  $\mathbf{f}$  по  $\boldsymbol{\theta}$ :

$$F_{ij} = \frac{\partial f_i}{\partial \theta_j}.$$

По сути дела мы здесь линеаризируем функцию  $\mathbf{f}$  в окрестности точки  $\boldsymbol{\theta}^t$ . Пусть  $\Delta\boldsymbol{\theta}^t$  — оценки ОМНК из этой вспомогательной регрессии:

$$\Delta\boldsymbol{\theta}^t = (\mathbf{F}(\boldsymbol{\theta}^t)^T \mathbf{F}(\boldsymbol{\theta}^t))^{-1} \mathbf{F}(\boldsymbol{\theta}^t)^T (\mathbf{Y} - \mathbf{f}(\boldsymbol{\theta}^t)).$$

Тогда следующее приближение метода Гаусса-Ньютона будет:

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \Delta\boldsymbol{\theta}^t.$$

Повторяем эти итерации пока метод не сойдется.

Последняя из регрессий Гаусса-Ньютона даст состоятельную оценку матрицы ковариаций оценок  $\hat{\boldsymbol{\theta}}$  ( $\hat{\sigma}^2 (\hat{\mathbf{F}}^T \hat{\mathbf{F}})^{-1}$ ), при условии, что верны обычные предположения: что  $\varepsilon_i$  независимо нормально распределены с нулевым мат. ожиданием и одинаковой дисперсией. Ясно, что можно, используя  $t$ - и  $F$ -статистики из этой вспомогательной регрессии, проверять различные гипотезы по принципу теста Вальда. Таким образом, регрессия Гаусса-Ньютона является искусственной регрессией.

## Оценивание регрессии с AR-ошибкой

AR(1) имеет вид:

$$\varepsilon_i = \rho \varepsilon_{i-1} + \xi_i.$$

“Инновации”  $\xi_i$  независимы и имеют одинаковую дисперсию  $\omega^2$ . Из этого следует, что  $\varepsilon_{i-1}$  и  $\xi_i$  независимы. Дисперсия  $\varepsilon_i$  находится из соотношения

$$V(\varepsilon_i) = V(\rho \varepsilon_{i-1}) + \omega^2.$$

$$\sigma^2 = \rho^2 \sigma^2 + \omega^2,$$

где  $\sigma^2 = V(\varepsilon_i)$ . Отсюда

$$\sigma^2 = \frac{\omega^2}{1 - \rho^2}.$$

Чтобы дисперсия не увеличивалась до бесконечности с ростом количества наблюдений, должно выполняться условие стационарности  $|\rho| < 1$ .

Найдем матрицу ковариаций ошибок  $\varepsilon_i$ . Рекуррентную формулу для  $\varepsilon_i$  можно развернуть следующим образом:

$$\varepsilon_i = \xi_i + \rho \xi_{i-1} + \rho^2 \xi_{i-2} + \dots + \rho^{\tau} \xi_{i-\tau}.$$

Отсюда  $\text{cov}(\varepsilon_i, \varepsilon_{i-\tau}) = \text{cov}(\rho^{\tau} \varepsilon_{i-\tau}, \varepsilon_{i-\tau}) = \rho^{\tau} \sigma^2$ .

Получаем следующую ковариационную матрицу

$$V(\sigma^2, \rho) = \sigma^2 \begin{bmatrix} 1 & \rho & \dots & \rho^{N-1} \\ \rho & \ddots & & \vdots \\ \vdots & & \ddots & \rho \\ \rho^{N-1} & \dots & \rho & 1 \end{bmatrix} = \sigma^2 W(\rho).$$

Предположим, что в линейной регрессионной модели  $Y_i = X_i \beta + \varepsilon_i$  ошибка порождается авторегрессионным процессом первого порядка. Рассмотрим два способа оценивания такой регрессионной модели.

### Нелинейная регрессия с пропущенным первым наблюдением

Подставим  $\varepsilon_i = Y_i - X_i \beta$  в уравнение авторегрессионного процесса:

$$Y_i - X_i \beta = \rho (Y_{i-1} - X_{i-1} \beta) + \xi_i.$$

Получим следующую нелинейную регрессионную модель:

$$Y_i = X_i \beta + \rho (Y_{i-1} - X_{i-1} \beta) + \xi_i.$$

Ошибки  $\xi_i$  независимые с одинаковой дисперсией и их ковариационная матрица равна  $\omega^2 I$ .

Для оценивания нелинейной регрессии можно использовать метод Гаусса-Ньютона. Поскольку  $Y_0$ ,  $X_0$  и  $\varepsilon_0$  неизвестны, то первое наблюдение не ис-

пользуют. С помощью вспомогательной регрессии метода Гаусса-Ньютона можно не только оценить модель, но и проверить гипотезу об отсутствии автокорреляции ( $\rho = 0$ ).

Если неизвестно, есть ли автокорреляция ошибок, лучше сначала получить оценки ОМНК и проверить гипотезу в этой точке (принцип LM-теста). Такой тест можно реализовать воспользовавшись той же регрессией Гаусса-Ньютона. В точке оценок ОМНК (когда  $\rho = 0$ ) матрица производных равна  $[X, e_{-1}]$ , т.е. к  $X$  добавляется столбец лагов остатков (где первое наблюдение равно нулю). Вспомогательная регрессия имеет вид

$$e = Xb + r e_{-1} + \text{ошибка}.$$

Проверяем гипотезу, что  $r = 0$ . Для этого можно использовать обычную  $t$ -статистику. Поскольку модель линейна, то это все равно, что тест на добавление переменной  $e_{-1}$  в исходную регрессию, так как можно заменить в левой части  $e$  на  $Y$ .

Этот тест и этот метод годятся даже тогда, когда в правой части стоят лаги зависимой переменной (DW-статистика в этом случае непригодна). Идею теста предложил Дарбин.

Описанный метод дает оценки МП при предположении, что ошибки распределены нормально, но не является точным методом МП, поскольку не учитывает распределение первого наблюдения.

### **Оценивание регрессии с AR(1)-ошибкой полным методом максимального правдоподобия**

Ковариационная матрица ошибок  $\varepsilon_i$  имеет вид:

$$V(\sigma^2, \rho) = \sigma^2 W(\rho) = \sigma^2 \begin{bmatrix} 1 & \rho & \dots & \rho^{N-1} \\ \rho & \ddots & & \vdots \\ \vdots & & \ddots & \rho \\ \rho^{N-1} & \dots & \rho & 1 \end{bmatrix}.$$

Дисперсии  $\varepsilon_i$  и  $\xi_i$  связаны между собой соотношением  $\sigma^2 = \frac{\omega^2}{1 - \rho^2}$ .

Можно проверить, что

$$W^{-1} = (1-\rho^2) \begin{bmatrix} 1 & -\rho & & 0 \\ -\rho & 1+\rho^2 & \ddots & 0 \\ & \ddots & \ddots & \ddots \\ 0 & & \ddots & 1+\rho^2 & -\rho \\ & & & -\rho & 1 \end{bmatrix}.$$

Применим следующее разложение Холецкого:  $TT^T = W^{-1}(1-\rho^2)$ , где

$$T = \begin{bmatrix} \sqrt{1-\rho^2} & -\rho & & 0 \\ & 1 & \ddots & \\ & & \ddots & -\rho \\ 0 & & & 1 \end{bmatrix}.$$

Определим переменные вспомогательной регрессии следующим образом:

$$Y^* = T^T Y, \quad X^* = T^T X.$$

При фиксированном  $\rho$  регрессия  $Y^*$  по  $X^*$  дает оценки максимума правдоподобия для  $\beta$  (это будет оценка МП только в том случае, если  $X$  не содержит лагов зависимой переменной !!!). Этот прием называется преобразованием Прэйса-Винстена (Prais-Winsten transformation). Распишем его более подробно:

$$\begin{aligned} Y_1^* &= \sqrt{1-\rho^2} Y_1, & X_{1j}^* &= \sqrt{1-\rho^2} X_{1j}, \\ Y_i^* &= Y_i - \rho Y_{i-1}, & X_{ij}^* &= X_i - \rho X_{i-1j} \quad \forall i > 1. \end{aligned}$$

Как и следовало ожидать, при  $i > 1$  преобразование совпадает с рассмотренным выше преобразованием при пропущенном первом наблюдении.

В данном случае формула для первого наблюдения отличается от формулы для прочих наблюдений. Поскольку в ней отсутствуют лаги, то ошибка будет равна  $\varepsilon_i$ , а не  $\xi_i$  и дисперсия для первого наблюдения будет равна  $\sigma^2 = \frac{\omega^2}{1-\rho^2}$ , а не  $\omega^2$ . Поэтому первое наблюдение домножается на  $\sqrt{1-\rho^2}$ , чтобы избавиться от гетероскедастичности.

Пусть  $\beta(\rho)$  — оценки коэффициентов из вспомогательной регрессии,  $e^*(\rho)$  — остатки из вспомогательной регрессии. Мы можем максимизировать по  $\rho$  концентрированную функцию правдоподобия:

$$\ell^c(\rho) = -\frac{1}{2} \ln |W(\rho)| - \frac{N}{2} \ln ((Y - X\beta(\rho))^T W(\rho)^{-1} (Y - X\beta(\rho))) +$$

$$+ \text{const} = \frac{1}{2} \ln(1-\rho^2) - \frac{N}{2} \ln(e^*(\rho)^T e^*(\rho)) + \text{const} \rightarrow \max_{\rho}.$$

Здесь мы воспользовались тем, что  $\mathbf{T}\mathbf{T}^T = \mathbf{W}^{-1}(1-\rho^2)$  и

$$\begin{aligned} \ln|\mathbf{W}| &= -\ln|\mathbf{W}^{-1}| = -N \ln(1-\rho^2) - \ln|\mathbf{T}\mathbf{T}^T| = \\ &= -N \ln(1-\rho^2) - 2 \ln|\mathbf{T}| = -N \ln(1-\rho^2) - 2 \ln(\sqrt{1-\rho^2}) = \\ &= -(N+1) \ln(1-\rho^2). \end{aligned}$$

Можно показать, что условие первого порядка максимума концентрированной функции правдоподобия представимо в виде кубического уравнения. Максимум находится как средний корень этого уравнения.

Удобно, что при этом оценка  $\rho$  всегда лежит в интервале стационарности  $(-1, 1)$ .

Информационная матрица, как и всегда в модели обобщенного метода наименьших квадратов, является блочно-диагональной. Приводим ее выборочный аналог без доказательства

$$\mathcal{I}^N(\hat{\boldsymbol{\theta}})^{-1} = \begin{bmatrix} \hat{\omega}^2 (\hat{\mathbf{X}}^{*\top} \hat{\mathbf{X}}^*)^{-1} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \frac{N}{1-\hat{\rho}^2} + \frac{3\hat{\rho}^2-1}{(1-\hat{\rho}^2)^2} & \frac{2\hat{\rho}}{\hat{\omega}(1-\hat{\rho})^2} \\ \mathbf{O} & \frac{2\hat{\rho}}{\hat{\omega}(1-\hat{\rho})^2} & \frac{2N}{\hat{\omega}^2} \end{bmatrix}.$$

Иногда первое наблюдение очень важно и добавляет много новой информации.

## Регрессия с МА-ошибкой

### Оценивание регрессия с МА(1)-процессом в ошибке полным методом максимального правдоподобия

Будем рассматривать регрессию  $Y = X\beta + \varepsilon$  с МА(1)-процессом в ошибке:

$$\varepsilon_i = \xi_i + \mu \xi_{i-1} \quad \xi_i \sim N(0, \omega^2)$$

$$\sigma^2 = \text{var}(\varepsilon_i) = (1 + \mu^2)\omega^2$$

Ковариационная матрица ошибок  $\varepsilon_i$  имеет вид  $V(\sigma^2, \mu) = \sigma^2 W(\mu)$ .

$$W = \begin{bmatrix} 1 + \mu^2 & \mu & & & \mathbf{O} \\ \mu & 1 + \mu^2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \mathbf{O} & \ddots & 1 + \mu^2 & \mu \\ & & & \mu & \mu & 1 + \mu^2 \end{bmatrix} = (1 + \mu^2)\mathbf{I} + \mu\Phi,$$

$$\text{где } \Phi = \begin{bmatrix} 0 & 1 & & & \mathbf{O} \\ 1 & 0 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \mathbf{O} & \ddots & 0 & 1 \\ & & & & 1 & 0 \end{bmatrix}.$$

Симметричную положительно определенную матрицу можно представить в виде  $W = H^T \Lambda H$ , где  $H$  — ортогональная матрица собственных векторов ( $H^T = H^{-1}$ ), а  $\Lambda$  — диагональная матрица, диагональ которой состоит из соответствующих собственных чисел. В данном случае, собственные вектора матрицы  $W$  совпадают с собственными векторами матрицы  $\Phi$ , и поэтому не зависят от  $\mu$ . Типичный элемент матрицы  $H$  равен

$$H_{kl} = \sin\left(\frac{kl\pi}{N+1}\right) \sqrt{\frac{2}{N+1}}.$$

Типичный диагональный элемент матрицы  $\Lambda$  (собственное число) равен

$$\lambda_k = \mu^2 + 2\mu \cos\left(\frac{k\pi}{N+1}\right) + 1.$$

Матрица  $W$  такова, что

$$W^{-1} = H^T \Lambda^{-1} H.$$

Несложно также найти определитель матрицы  $W$ :

$$|W| = \frac{1 - \mu^{2N+2}}{1 - \mu^2}.$$

Обозначим  $Y^* = DHY$ ,  $X^* = DHX$ , где  $D = \Lambda^{-1/2}$  — диагональная матрица.

Пусть  $e^*(\mu)$  — остатки из вспомогательной регрессии,  $\beta(\mu)$  — оценки коэффициентов из этой регрессии. Тогда

$$(Y - X\beta(\mu))^T W(\mu)^{-1} (Y - X\beta(\mu)) = e^*(\mu)^T e^*(\mu).$$

Концентрированная функция правдоподобия после исключения параметров  $\sigma^2$  и  $\beta$  приобретет вид

$$\ell^c(\mu) = -\frac{1}{2} (\ln(1 - \mu^{2N+2}) - \ln(1 - \mu^2)) - \frac{N}{2} \ln(e^*(\mu)^T e^*(\mu)) + \text{const}.$$

Остается с помощью одномерного поиска максимизировать концентрированную функцию правдоподобия по  $\mu$  на отрезке  $[-1, 1]$ . Максимум функции правдоподобия может с ненулевой вероятностью достигаться при  $\mu = 1$  или  $\mu = -1$ .

Можно предложить и другую вспомогательную регрессию, которая применима и в общем случае  $MA(l)$ -процесса. Обозначим  $\eta = \xi_0$ . Тогда модель можно преобразовать к виду

$$0 = -\eta + \xi_0,$$

$$Y_1 = X_1\beta + \mu\eta + \xi_1,$$

$$Y_2 - \mu Y_1 = (X_2 - \mu X_1)\beta - \mu^2\eta + \xi_2,$$

и так далее для  $i = 3, \dots, N$ .

Более компактно это можно записать как уравнение регрессии:

$$Y = X\beta + Z\eta + \xi.$$

Здесь  $Y$ ,  $X$  и  $Z$  имеют по  $N+1$  наблюдению и вычисляются по рекуррентным формулам:

$$Y_i = Y_i - \mu Y_{i-1}, \quad Y_0 = 0,$$

$$X_i = X_i - \mu X_{i-1}, \quad X_0 = \mathbf{0},$$

$$Z_i = -\mu Z_{i-1}, \quad Z_0 = -1.$$

Пусть  $\xi(\mu)$  — остатки из вспомогательной регрессии,  $\beta(\mu)$  — оценки коэффициентов  $\beta$  из этой регрессии. Тогда можно показать, что  $(Y -$

$X\beta(\mu)^T W(\mu)^{-1}(Y - X\beta(\mu)) = \xi(\mu)^T \xi(\mu)$ . Соответственно, концентрированная функция правдоподобия равна

$$\ell^c(\mu) = -\frac{1}{2} (\ln(1-\mu^{2N+2}) - \ln(1-\mu^2)) - \frac{N}{2} \ln(\xi(\mu)^T \xi(\mu)) + \text{const.}$$

### Оценивание регрессии с МА-ошибкой нелинейным МНК

Как и в случае регрессии с AR-процессом, если пренебречь первыми наблюдениями, можно получить оценку, которая асимптотически эквивалентна точной оценке максимального правдоподобия. В данном случае удобно считать, что довыборочные ошибки  $\xi_i$  ( $i < 1$ ) равны нулю. При этом из функции правдоподобия исчезает мешающий член  $-1/2 (\ln(1-\mu^{2N+2}) - \ln(1-\mu^2))$ , и модель сводится к нелинейной регрессии, которую можно оценить с помощью метода Гаусса-Ньютона. Требуется минимизировать сумму квадратов остатков

$$\sum_{i=1}^N \xi_i^2(\beta, \mu) \rightarrow \min.$$

Остатки вычисляются рекуррентно по формуле

$$\xi_i(\beta, \mu) = Y_i - X_i \beta - \mu \xi_{i-1}(\beta, \mu) \quad (\xi_0(\beta, \mu) = 0).$$

Производные функции  $\xi_i(\beta, \mu)$ , необходимые для использования метода Гаусса-Ньютона также находятся рекуррентно:

$$\begin{aligned} \frac{\partial \xi_i}{\partial \beta} &= -X_i - \mu \frac{\partial \xi_{i-1}}{\partial \beta} \quad \left( \frac{\partial \xi_0}{\partial \beta} = 0 \right). \\ \frac{\partial \xi_i}{\partial \mu} &= -\mu \frac{\partial \xi_{i-1}}{\partial \mu} - \xi_{i-1} \quad \left( \frac{\partial \xi_0}{\partial \mu} = 0 \right). \end{aligned}$$

## Регрессия с ARCH-процессом в ошибке

Часто в эконометрических моделях остатки становятся то большими на какой-то период, то не очень большими, и так далее и в этом нет определенной закономерности. Особенно это относится к моделям финансовых рынков. Даже если безусловная дисперсия ошибок постоянна, условная дисперсия может быть подвержена случайным колебаниям. Условные прогнозы дисперсии могут иметь практическое значение. Владелец активов интересуется, как получить прогноз риска на ряд следующих периодов, если известна информация за текущий и предшествующие периоды.

Для моделирования таких процессов используется понятие условной авторегрессионной гетероскедастичности (ARCH, autoregressive conditional heteroskedasticity). Ввел это понятие Энгл (Engle, 1982). Основная идея состоит в том, что дисперсия ошибки в момент  $i$  зависит от величины квадрата ошибки в предшествующие периоды времени.

Процесс ARCH( $p$ ) имеет следующий вид:

$$\sigma_i^2 \equiv E(\varepsilon_i^2 | \Omega_i) = \mu + \gamma_1 \varepsilon_{i-1}^2 + \gamma_2 \varepsilon_{i-2}^2 + \dots + \gamma_p \varepsilon_{i-p}^2.$$

$\Omega_i$  обозначает информацию, которая имеется к моменту  $i$ , т. е. все наблюдаемые переменные в момент  $i-1$  и ранее. В частности, для ARCH(1)

$$\sigma_i^2 = \mu + \gamma_1 \varepsilon_{i-1}^2.$$

Предполагают, что  $\mu > 0$ ,  $\gamma_k \geq 0$ , чтобы дисперсия не могла оказаться отрицательной.

В случае, когда  $\varepsilon_i$  имеют нормальное распределение, очередное значение задается формулой

$$\varepsilon_i = \sigma_i \xi_i, \quad \text{где } \xi_i \sim \text{NID}(0,1).$$

или 
$$\varepsilon_i = \xi_i \sqrt{\mu + \gamma_1 \varepsilon_{i-1}^2 + \gamma_2 \varepsilon_{i-2}^2 + \dots + \gamma_p \varepsilon_{i-p}^2}.$$

Такой процесс имеет большое сходство с авторегрессионным.

Обозначим безусловную дисперсию через  $\sigma^2$ . Тогда, если взять ожидания от обеих частей в формуле для дисперсии, получим

$$\sigma^2 = \mu + \gamma_1 \sigma^2 + \gamma_2 \sigma^2 + \dots + \gamma_p \sigma^2.$$

Здесь мы воспользовались формулой полного математического ожидания:

$$E(\sigma_i^2) = E(E(\varepsilon_i^2 | \Omega_i)) = E(\varepsilon_i^2) = \sigma^2.$$

Получаем выражение для безусловной дисперсии

$$\sigma^2 = \frac{\mu}{1 - \sum_{k=1}^p \gamma_k}.$$

Отсюда видно, что для того, чтобы ARCH-процесс был стационарным необходимо, чтобы  $\sum_{k=1}^p \gamma_k < 1$ .

На практике структура лага  $\gamma_1, \dots, \gamma_p$  неизвестна и ее приходится оценивать. Если  $p$  большое, то приходится оценивать слишком много коэффициентов. В таком случае можно наложить априорные ограничения. Энгл, например, взял

$$\gamma_1 = 0.4\gamma^*, \quad \gamma_2 = 0.3\gamma^*, \quad \gamma_3 = 0.2\gamma^*, \quad \gamma_4 = 0.1\gamma^*,$$

поэтому вместо четырех коэффициентов, ему надо было оценить только один  $\gamma^*$ .

Если воспользоваться бесконечным геометрическим лагом, то можно преобразовать модель к виду

$$\sigma^2 = \mu + \gamma_1 \varepsilon_{i-1}^2 + \delta_1 \sigma_{i-1}^2.$$

Эта модель называется GARCH(1,1). Как ARCH похожа AR на, так GARCH похожа на ARMA. Модель GARCH предложена Боллерселевом (Bollerslev, 1986).

GARCH( $p, q$ ) имеет вид:

$$\begin{aligned} \sigma_i^2 &= \mu + \gamma_1 \varepsilon_{i-1}^2 + \dots + \gamma_p \varepsilon_{i-p}^2 + \delta_1 \sigma_{i-1}^2 + \dots + \delta_q \sigma_{i-q}^2 = \\ &= \mu + \sum_{k=1}^p \gamma_k \varepsilon_{i-k}^2 + \sum_{k=1}^q \delta_k \sigma_{i-k}^2. \end{aligned}$$

Безусловная дисперсия равна

$$\sigma^2 = \frac{\mu}{1 - \sum_{k=1}^p \gamma_k - \sum_{k=1}^q \delta_k}.$$

Регрессионная модель с нормально распределенной GARCH-ошибкой имеет вид:

$$Y_i = X_i \beta + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_i^2),$$

где  $\sigma_i^2$  вычисляется по данной выше формуле. К сожалению, не существует простого способа оценивания регрессии с GARCH-процессом в ошибке.

Логарифмическая функция правдоподобия для одного наблюдения (с точностью до константы) равна

$$\ell_i = -\frac{1}{2} \ln \sigma_i^2 - \frac{1}{2} \frac{e_i^2}{\sigma_i^2} = -\frac{1}{2} \ln \sigma_i^2 - \frac{1}{2} \frac{1}{\sigma_i^2} (Y_i - \mathbf{X}_i \boldsymbol{\beta})^2,$$

для всего вектора наблюдений

$$\begin{aligned} \ell &= \sum_i \ell_i = -\frac{1}{2} \sum_i \ln \sigma_i^2 - \frac{1}{2} \sum_i \frac{e_i^2}{\sigma_i^2} = \\ &= -\frac{1}{2} \sum_i \ln \sigma_i^2 - \frac{1}{2} \sum_i \frac{1}{\sigma_i^2} (Y_i - \mathbf{X}_i \boldsymbol{\beta})^2, \end{aligned}$$

где  $\sigma_i^2$  вычисляется рекуррентно:

$$\sigma_i^2(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\delta}) = \mu + \sum_{k=1}^p \gamma_k e_{i-k}^2(\boldsymbol{\beta}) + \sum_{k=1}^q \delta_k \sigma_{i-k}^2(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\delta}).$$

Обозначим  $\boldsymbol{\alpha} = (\mu, \gamma_1, \dots, \gamma_p, \delta_1, \dots, \delta_q)^\top$ ,  $\mathbf{z}_i = (1, \varepsilon_{i-1}^2, \dots, \varepsilon_{i-p}^2, \sigma_{i-1}^2, \dots, \sigma_{i-p}^2)^\top$ .

Тогда  $\sigma_i^2 = \boldsymbol{\alpha}^\top \mathbf{z}_i$ .

Вклад в градиент отдельного наблюдения:

$$\frac{\partial \ell_i}{\partial \boldsymbol{\beta}} = \frac{1}{2} \frac{1}{\sigma_i^2} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\beta}} \left( \frac{e_i^2}{\sigma_i^2} - 1 \right) + \frac{1}{\sigma_i^2} e_i \mathbf{X}_i,$$

$$\frac{\partial \ell_i}{\partial \boldsymbol{\alpha}} = \frac{1}{2} \frac{1}{\sigma_i^2} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\alpha}} \left( \frac{e_i^2}{\sigma_i^2} - 1 \right).$$

Производные условной дисперсии вычисляются рекуррентно:

$$\frac{\partial \sigma_i^2}{\partial \boldsymbol{\beta}} = -2 \sum_{k=1}^p \gamma_k \mathbf{X}_{i-k} e_{i-k} + \sum_{k=1}^q \delta_k \frac{\partial \sigma_{i-k}^2}{\partial \boldsymbol{\beta}},$$

$$\frac{\partial \sigma_i^2}{\partial \boldsymbol{\alpha}} = \mathbf{z}_i^\top + \sum_{k=1}^q \delta_k \frac{\partial \sigma_{i-k}^2}{\partial \boldsymbol{\alpha}}.$$

Приведенные формулы уже позволяют применить метод ВНИН (OPG), как предложил Боллерслев. Однако, метод, как обычно, работает плохо в смысле сходимости и точности оценивания ковариационной матрицы оценок МП. Другие методы, — Ньютона, удвоенной регрессии, — работают гораздо лучше. Рассмотрим метод, использующий оценку информационной матри-

цы, в которой ожидания берутся только частично — только условные ожидания.

Чтобы найти оценку информационной матрицы, возьмем в точке истинных параметров математическое ожидание условное по информации, имеющейся на момент  $i$  ( $\Omega_i$ ), от матрицы внешнего произведения градиента  $i$ -го наблюдения. Поскольку ожидание условное, то единственной случайной компонентой будет ошибка (в точке истинных параметров остатки равны ошибкам). При этом пользуемся тем, что

$$E\left(\frac{\varepsilon_i^2}{\sigma_i^2} - 1 \mid \Omega_i\right) = 0, \quad E\left(\left(\frac{\varepsilon_i^2}{\sigma_i^2} - 1\right)^2 \mid \Omega_i\right) = 2 \quad \text{и} \quad E(\varepsilon_i^2 \mid \Omega_i) = \sigma_i^2.$$

Отсюда

$$E\left(\frac{\partial l_i}{\partial \boldsymbol{\beta}^\top} \frac{\partial l_i}{\partial \boldsymbol{\beta}} \mid \Omega_i\right) = \frac{1}{2} \frac{1}{(\sigma_i^2)^2} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\beta}^\top} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\beta}} + \frac{1}{\sigma_i^2} \mathbf{X}_i^\top \mathbf{X}_i,$$

$$E\left(\frac{\partial l_i}{\partial \boldsymbol{\alpha}^\top} \frac{\partial l_i}{\partial \boldsymbol{\alpha}} \mid \Omega_i\right) = -\frac{1}{2} \frac{1}{(\sigma_i^2)^2} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\alpha}^\top} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\alpha}},$$

$$E\left(\frac{\partial l_i}{\partial \boldsymbol{\beta}^\top} \frac{\partial l_i}{\partial \boldsymbol{\alpha}} \mid \Omega_i\right) = -\frac{1}{2} \frac{1}{(\sigma_i^2)^2} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\beta}^\top} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\alpha}}.$$

Информационная матрица равна безусловному ожиданию суммы условных мат. ожиданий гессианов со знаком минус:

$$\mathcal{I}^N(\boldsymbol{\theta}) = E\left(\sum_i E\left(\frac{\partial l_i}{\partial \boldsymbol{\theta}} \frac{\partial l_i}{\partial \boldsymbol{\theta}^\top} \mid \Omega_i\right)\right)$$

$$\mathcal{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}^N = \sum_i E\left(\frac{1}{2} \frac{1}{(\sigma_i^2)^2} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\beta}} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\beta}^\top} + \frac{1}{\sigma_i^2} \mathbf{X}_i \mathbf{X}_i^\top\right),$$

$$\mathcal{I}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^N = \sum_i E\left(\frac{1}{2} \frac{1}{(\sigma_i^2)^2} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\alpha}} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\alpha}^\top}\right), \quad \mathcal{I}_{\boldsymbol{\alpha}\boldsymbol{\beta}}^N = \sum_i E\left(\frac{1}{2} \frac{1}{(\sigma_i^2)^2} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\alpha}} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\beta}^\top}\right).$$

Можно показать, хотя доказательство достаточно громоздкое, что  $\mathcal{I}_{\boldsymbol{\alpha}\boldsymbol{\beta}}^N = \mathbf{O}$ . Таким образом, информационная матрица является блочно-диагональной между коэффициентами регрессии и параметрами GARCH-процесса.

Информационную матрицу можно состоятельно оценить матрицей:

$$\hat{\mathcal{I}}^N = \begin{bmatrix} \hat{\mathcal{I}}_{\beta\beta}^N & \mathbf{O} \\ \mathbf{O} & \hat{\mathcal{I}}_{\alpha\alpha}^N \end{bmatrix},$$

где

$$\hat{\mathcal{I}}_{\beta\beta}^N = \sum_i \frac{1}{2} \frac{1}{(\sigma_i^2)^2} \frac{\partial \sigma_i^2}{\partial \beta} \frac{\partial \sigma_i^2}{\partial \beta^\top} + \frac{1}{\sigma_i^2} \mathbf{X}_i \mathbf{X}_i^\top \text{ и } \hat{\mathcal{I}}_{\alpha\alpha}^N = \sum_i \frac{1}{2} \frac{1}{(\sigma_i^2)^2} \frac{\partial \sigma_i^2}{\partial \alpha} \frac{\partial \sigma_i^2}{\partial \alpha^\top}.$$

Для нахождения оценок максимального правдоподобия можно применить обычный алгоритм:

$$\begin{aligned} \beta^{t+1} &= \beta^t + \lambda_\beta (\hat{\mathcal{I}}_{\beta\beta}^N)^{-1} \mathbf{g}_\beta(\theta^t), \\ \alpha^{t+1} &= \alpha^t + \lambda_\alpha (\hat{\mathcal{I}}_{\alpha\alpha}^N)^{-1} \mathbf{g}_\alpha(\theta^t). \end{aligned}$$

Прежде чем оценивать GARCH-регрессию, имеет смысл проверить наличие GARCH-процесса в ошибке с помощью теста множителя Лагранжа в точке оценок ОМНК. Используем блочно-диагональность информационной матрицы:

$$\text{LM} = \tilde{\mathbf{g}}_\alpha^\top (\tilde{\mathcal{I}}_{\alpha\alpha}^N)^{-1} \tilde{\mathbf{g}}_\alpha$$

При нулевой гипотезе об отсутствии GARCH-процесса  $\sigma_i^2 = \mu = \sigma^2$ , и статистику можно упростить. Найдем эту статистику только в частном случае ARCH модели:

$$\frac{\partial \sigma_i^2}{\partial \alpha} = \mathbf{z}_i^\top \Rightarrow \frac{\partial \ell_i}{\partial \alpha} = \frac{1}{2} \frac{1}{\sigma^2} \mathbf{z}_i^\top \left( \frac{e_i^2}{\sigma^2} - 1 \right), \quad \tilde{\mathcal{I}}_{\alpha\alpha}^N = \sum_i \frac{1}{2} \frac{1}{(\sigma^2)^2} \mathbf{z}_i^\top \mathbf{z}_i.$$

Пусть  $\boldsymbol{\mu}$  — вектор-столбец, состоящий из  $\frac{e_i^2}{\sigma^2} - 1$ ,  $\mathbf{Z}$  — матрица, строками которой являются  $\mathbf{z}_i^\top = (1, e_{i-1}^2, \dots, e_{i-p}^2)$ . Тогда

$$\tilde{\mathbf{g}}_\alpha^\top = \frac{\partial \ell}{\partial \alpha} = \frac{1}{2} \frac{1}{\sigma^2} \boldsymbol{\mu}^\top \mathbf{Z}, \quad \tilde{\mathcal{I}}_{\alpha\alpha}^N = \frac{1}{2} \frac{1}{(\sigma^2)^2} \mathbf{Z}^\top \mathbf{Z}.$$

$$\text{LM} = \frac{1}{2} \boldsymbol{\mu}^\top \mathbf{Z} (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \boldsymbol{\mu} \overset{a}{\sim} \chi^2(p).$$

Энгл предложил несколько изменить форму статистики. Для этого используется то, что поскольку ошибки распределены нормально, то

$$\text{Plim } \boldsymbol{\mu}^\top \boldsymbol{\mu} / N = 2.$$

Асимптотически эквивалентной статистикой будет

$$LM^* = \frac{N}{\mathbf{\mu}^\top \mathbf{\mu}} \mathbf{\mu}^\top \mathbf{Z} (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{\mu}^a \chi^2(p).$$

Эта статистика равняется  $NR^2$ , где  $R^2$  — коэффициент детерминации в регрессии квадратов остатков по  $\mathbf{Z}$ , то есть по константе и  $p$  лагам квадратов остатков.

## Якобиан преобразования плотности распределения в функции правдоподобия

### Функция правдоподобия модели типа $\varepsilon = f(Y, \theta_1)$

Рассмотрим модель по отношению к которой регрессия является частным случаем:

$$\varepsilon = f(Y, \theta_1).$$

Здесь  $Y$  — зависимая переменная,  $\varepsilon$  — ошибка, причем  $Y$  и  $\varepsilon$  — вектора-столбцы одинаковой размерности. “Независимые переменные” (регрессоры)  $X$  неявно содержатся в функции  $f(\cdot)$ .  $\theta_1$  — неизвестные параметры. Обозначаем их  $\theta_1$ , а не  $\theta$ , потому что распределение ошибок  $\varepsilon$  само может зависеть от вектора неизвестных параметров ( $\theta_2$ ), так что

$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$$

В частном случае линейной регрессии

$$f(Y, \theta_1) = Y - X\beta, \quad \theta_1 = \beta, \quad \theta_2 = \sigma^2.$$

Как правило, при построении эконометрической модели делают предположение о распределении ошибки, а уже из распределения ошибки выводят распределение зависимой переменной. Таким образом, задача состоит в том, чтобы из плотности распределения  $\varepsilon$  получить плотность распределения  $Y$  (если мы имеем дело с непрерывным распределением).

Плотности распределения связаны между собой соотношением:

$$p_Y(Y, \theta) = p_\varepsilon(f(Y, \theta_1), \theta_2) \text{ abs } |\mathbf{J}(\theta_1)|,$$

где  $\mathbf{J}(\theta_1)$  — матрица Якоби (якобиан), соответствующий преобразованию  $Y$  в  $\varepsilon$ :

$$\mathbf{J}(\theta_1) = \frac{\partial \mathbf{f}}{\partial Y} = \left\{ \frac{\partial f_i}{\partial Y_j} \right\}_{il}$$

— матрица первых производных  $f$  по  $Y$ . В выражении для плотности здесь стоит модуль определителя якобиана.

Функция правдоподобия — это по определению плотность распределения  $Y$ . Таким образом, логарифмическая функция правдоподобия равна

$$\ell = \ln p_\varepsilon(f(Y, \theta_1), \theta_2) + \ln \text{ abs } |\mathbf{J}(\theta_1)|.$$

Будем второе слагаемое здесь называть якобианным членом. Якобианный член уже присутствовал в логарифмических функциях правдоподобия, которые мы рассматривали выше (см. напр. регрессии с автокоррелированным

ными ошибками). В модели с AR(1)-ошибкой  $\varepsilon_i = \rho \varepsilon_{i-1} + \xi_i$ , где  $\varepsilon_i = Y_i - X_i \beta$ . Выразим  $\xi$  через  $Y$ :

$$f_i(Y, \theta_1) = \xi_i = (Y_i - X_i \beta) - \rho (Y_{i-1} - X_{i-1} \beta), \quad i=1, \dots, N.$$

$$f_1(Y, \theta_1) = \sqrt{1 - \rho^2} (Y_1 - X_1 \beta).$$

Здесь  $\theta_1 = \begin{pmatrix} \beta \\ \rho \end{pmatrix}$ .  $f_i(Y, \theta_1)$  определена таким образом, чтобы все элементы  $f$  имели одинаковую дисперсию. Для этой модели

$$\mathbf{J}(\theta_1) = \frac{\partial f}{\partial Y} = \begin{bmatrix} \sqrt{1-\rho^2} & & & \mathbf{O} \\ -\rho & 1 & & \\ & \ddots & \ddots & \\ \mathbf{O} & & -\rho & 1 \end{bmatrix}, \quad \text{abs } |\mathbf{J}(\theta_1)| = \sqrt{1 - \rho^2}.$$

### Преобразование зависимой переменной. Модель Бокса-Кокса

В частном случае рассмотренной модели  $\theta_1$  состоит из  $\beta$  и  $\delta$  и

$$f_i(Y_i, \theta_1) = h_i(Y_i, \delta) - X_i(\delta) \beta.$$

Модель является квазирегрессионной. Здесь  $X(\delta)$  — матрица “регрессоров”,  $\beta$  — вектор регрессионных коэффициентов. Якобиан  $\mathbf{J} = \frac{\partial f}{\partial Y} = \frac{\partial h}{\partial Y}$  зависит только от параметров  $\delta$  и является диагональной матрицей.

Такая модель возникает из регрессии, если применить к зависимой переменной преобразование, зависящее от оцениваемых параметров.

Пусть ошибки нормально распределены  $\varepsilon_i \sim N(0, \sigma^2)$  и некоррелированы.

$$p_\varepsilon(z) = (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{1}{2\sigma^2} z^\top z\right).$$

Логарифмическая функция правдоподобия для такой модели:

$$\begin{aligned} \ell &= \ln p_\varepsilon(f(Y, \theta_1), \theta_2) + \ln \text{abs } |\mathbf{J}(\theta_1)| = \\ &= -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} f^\top f + \ln \text{abs } \left| \frac{\partial f}{\partial Y} \right| = \\ &= -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_i (h_i(Y_i, \delta) - X_i(\delta) \beta)^2 + \sum_i \ln \text{abs} \left( \frac{\partial h_i}{\partial Y_i} \right). \end{aligned}$$

Самое популярное преобразование зависимой переменной — это **преобразование Бокса-Кокса**:

$$h_i(Y_i, \delta) = \frac{Y_i^\delta - 1}{\delta}.$$

В общем случае его можно применять, только если все  $Y_i$  положительны.

Если  $\delta \rightarrow 0$ , то  $\frac{Y_i^\delta - 1}{\delta} \rightarrow \ln Y_i$ , поэтому берут  $h(Y, 0) = \ln Y$ .

Таким образом, имеем следующую простейшую модель Бокса-Кокса<sup>3</sup>:

$$\frac{Y_i^\delta - 1}{\delta} = \mathbf{X}_i \boldsymbol{\beta} + \varepsilon_i.$$

Здесь регрессоры  $\mathbf{X}$  детерминированы и не зависят от неизвестных параметров.

Якобиан равен

$$\mathbf{J} = \begin{bmatrix} Y_1^{\delta-1} & & \mathbf{O} \\ & \ddots & \\ \mathbf{O} & & Y_N^{\delta-1} \end{bmatrix}.$$

Функция правдоподобия для модели Бокса-Кокса равна

$$\ell = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_i \left( \frac{Y_i^\delta - 1}{\delta} - \mathbf{X}_i \boldsymbol{\beta} \right)^2 - (1 - \delta) \sum_i \ln Y_i.$$

Концентрируя функцию правдоподобия по  $\sigma^2$ , получим

$$\ell^c = -\frac{N}{2} \ln(2\pi \frac{\text{RSS}}{N}) - \frac{N}{2} - (1 - \delta) \sum_i \ln Y_i.$$

Обозначим  $\bar{Y}$  среднее геометрическое  $Y_i$ :

$$\bar{Y} = (\prod_i Y_i)^{1/N}.$$

Тогда  $\ell^c = -\frac{N}{2} \ln(\text{RSS}) - N(1 - \delta) \ln \bar{Y} + \text{const}$ .

Максимизация  $\ell^c$  эквивалентна минимизации следующей суммы квадратов:

$$\sum_i (\bar{Y}^{1-\delta} \left( \frac{Y_i^\delta - 1}{\delta} - \mathbf{X}_i \boldsymbol{\beta} \right))^2.$$

Можно предложить два метода оценивания.

Первый метод заключается в одномерной минимизации по  $\delta$ , поскольку при фиксированном  $\delta$  задача сводится к ОМНК. Строим регрессию  $\bar{Y}^{1-\delta} (Y_i^\delta - 1)/\delta$  по  $\bar{Y}^{1-\delta} \mathbf{X}_i$ .

---

<sup>3</sup> В моделях этого типа переменные в правой части также могут подвергаться преобразованию Бокса-Кокса.

Второй метод заключается в использовании нелинейного МНК, в котором зависимой переменной является вектор, состоящий из нулей, а в правой части стоит  $\bar{Y}^{1-\delta} ((Y_i^\delta - 1)/\delta - X_i \beta)$ .

В обоих случаях мы найдем МП-оценки, но не найдем состоятельную оценку матрицы ковариаций оценок (ковариационные матрицы из этих вспомогательных регрессий не годятся).

## Тест на нормальность

Задача этого параграфа — получить статистику множителя Лагранжа, которая позволила бы проверять гипотезу о том, что ошибки в регрессии распределены нормально. Идея состоит в том, чтобы рассмотреть модель с ошибкой из некоторого семейства непрерывных распределений, так чтобы нормальное распределение было частным случаем. Удобно взять, например, пирсоновское семейство распределений.

Плотность распределения (с нулевым мат. ожиданием) из пирсоновского семейства задается экспонентой функции

$$\psi(\varepsilon, c) = \int \frac{c_1 - t}{c_2 t^2 - c_1 t + c_0} dt.$$

Поскольку интеграл плотности распределения должен быть равен 1, то эту функцию следует пронормировать:

$$p_\varepsilon(u) = \frac{\exp \psi(u)}{\int_{-\infty}^{+\infty} \exp \psi(t) dt}.$$

Нулевая гипотеза (“нормальность”) заключается в том, что ошибки в линейной регрессии  $Y$  по  $X$  распределены нормально. Нормальное распределение является пирсоновским распределением с параметрами  $c_1 = 0$ ,  $c_2 = 0$ :

$$H_0: c_1 = 0, c_2 = 0 \Rightarrow \varepsilon \sim N(0, c_0) \text{ (при } c_0 = \sigma^2 \text{)}.$$

Логарифмическая функция правдоподобия есть логарифм плотности распределения. Для  $i$ -го наблюдения:

$$l_i = \psi(Y_i - X_i \beta) - \ln \int_{-\infty}^{+\infty} \exp \psi(t) dt.$$

Найдем вклад в градиент  $i$ -го наблюдения при выполнении нулевой гипотезы.

$$\frac{\partial l_i}{\partial \beta} = - \frac{\partial \psi}{\partial \varepsilon} X_i = - \frac{c_1 - e_i}{c_2 e_i^2 - c_1 e_i + c_0} X_i.$$

$$\left. \frac{\partial l_i}{\partial \beta} \right|_{H_0} = \frac{\varepsilon_i}{c_0} X_i = \frac{1}{\sigma^2} \varepsilon_i X_i.$$

Производные по параметрам  $c_k$  пирсоновского распределения равны

$$\begin{aligned} \frac{\partial \ell_i}{\partial c_k} &= \frac{\partial \Psi}{\partial c_k} - \frac{\int_{-\infty}^{+\infty} \frac{\partial \Psi(t)}{\partial c_k} \exp \Psi(t) dt}{\int_{-\infty}^{+\infty} \exp \Psi(t) dt} = \frac{\partial \Psi}{\partial c_k} - \int_{-\infty}^{+\infty} \frac{\partial \Psi(t)}{\partial c_k} p_{\varepsilon}(t) dt = \\ &= \frac{\partial \Psi}{\partial c_k} - E\left(\frac{\partial \Psi}{\partial c_k}\right) \quad (k = 0, 1, 2). \end{aligned}$$

Чтобы их вычислить, нужно вычислить производные функции  $\Psi(\cdot)$  по  $c_k$  ( $k = 0, 1, 2$ ). Достаточно найти их при нулевой гипотезе:

$$\left. \frac{\partial \Psi(u)}{\partial c_0} \right|_{H_0} = \frac{1}{c_0^2} \int^u t dt = \frac{u^2}{2\sigma^4}.$$

$$\left. \frac{\partial \Psi(u)}{\partial c_1} \right|_{H_0} = \frac{1}{c_0} \int^u dt - \frac{1}{c_0^2} \int^u t^2 dt = \frac{u}{\sigma^2} - \frac{u^3}{3\sigma^4}.$$

$$\left. \frac{\partial \Psi(u)}{\partial c_2} \right|_{H_0} = \frac{1}{c_0^2} \int^u t^3 dt = \frac{u^4}{4\sigma^4}.$$

Математические ожидания этих производных как функций от  $\varepsilon_i$  равны

$$E\left(\left. \frac{\partial \Psi(\varepsilon_i)}{\partial c_0} \right|_{H_0}\right) = E\left(\frac{\varepsilon_i^2}{2\sigma^4}\right) = \frac{1}{2\sigma^2},$$

$$E\left(\left. \frac{\partial \Psi(\varepsilon_i)}{\partial c_1} \right|_{H_0}\right) = E\left(\frac{\varepsilon_i}{\sigma^2} - \frac{\varepsilon_i^3}{3\sigma^4}\right) = 0,$$

$$E\left(\left. \frac{\partial \Psi(\varepsilon_i)}{\partial c_2} \right|_{H_0}\right) = E\left(\frac{\varepsilon_i^4}{4\sigma^4}\right) = \frac{3}{4}.$$

Подставим найденные выражения в градиент логарифмической функции правдоподобия, введя обозначение  $\tilde{\varepsilon}_i = \varepsilon_i/\sigma$ :

$$G_{0i}^0 = \left. \frac{\partial \ell_i}{\partial c_0} \right|_{H_0} = \frac{1}{2\sigma^4} (\varepsilon_i^2 - \sigma^2) = \frac{1}{2\sigma^2} (\tilde{\varepsilon}_i^2 - 1),$$

$$G_{0i}^1 = \left. \frac{\partial \ell_i}{\partial c_1} \right|_{H_0} = \frac{\varepsilon_i}{\sigma^2} - \frac{\varepsilon_i^3}{3\sigma^4} = \frac{1}{3\sigma} (3\tilde{\varepsilon}_i - \tilde{\varepsilon}_i^3),$$

$$G_{0i}^2 = \left. \frac{\partial \ell_i}{\partial c_2} \right|_{H_0} = \frac{\varepsilon_i^4}{4\sigma^4} - \frac{3}{4} = \frac{1}{4} (\tilde{\varepsilon}_i^4 - 3).$$

В тех же обозначениях

$$G_{0i}^{\mathbf{B}} = \left. \frac{\partial \ell_i}{\partial \mathbf{B}} \right|_{H_0} = \frac{1}{\sigma} \tilde{\varepsilon}_i \mathbf{X}_i.$$

Найдем информационную матрицу, учитывая, что моменты стандартного нормального распределения ( $\eta \sim N(0,1)$ ) равны

$$E(\eta^k) = \begin{cases} 0, & k \text{ — нечетное} \\ 1 \cdot 3 \cdot \dots \cdot (k-1), & k \text{ — четное} \end{cases},$$

$$E(\eta^4) = 3, \quad E(\eta^6) = 15, \quad E(\eta^8) = 105.$$

Информационная матрица для  $i$ -го наблюдения:

$$E(G_{0i}^{\beta \top} G_{0i}^{\beta}) = \frac{1}{\sigma^2} \mathbf{X}_i^{\top} \mathbf{X}_i E\tilde{\varepsilon}_i^2 = \frac{1}{\sigma^2} \mathbf{X}_i^{\top} \mathbf{X}_i,$$

$$E(G_{0i}^0 G_{0i}^{\beta}) = \mathbf{0}^{\top}, \quad E(G_{0i}^2 G_{0i}^{\beta}) = \mathbf{0}^{\top},$$

$$E(G_{0i}^1 G_{0i}^{\beta}) = \frac{1}{3\sigma^2} (3E\tilde{\varepsilon}_i^2 - E\tilde{\varepsilon}_i^4) \mathbf{X}_i = \frac{1}{3\sigma^2} (3 - 3) \mathbf{X}_i = \mathbf{0}^{\top},$$

$$E((G_{0i}^0)^2) = \frac{1}{4\sigma^4} (E\tilde{\varepsilon}_i^4 - 2E\tilde{\varepsilon}_i^2 + 1) = \frac{1}{4\sigma^4} (3 - 2 + 1) = \frac{1}{2\sigma^4},$$

$$E(G_{0i}^0 G_{0i}^1) = 0, \quad E(G_{0i}^1 G_{0i}^2) = 0,$$

$$E((G_{0i}^1)^2) = \frac{1}{9\sigma^2} (E\tilde{\varepsilon}_i^6 - 6E\tilde{\varepsilon}_i^4 + 9E\tilde{\varepsilon}_i^2) = \frac{1}{9\sigma^2} (15 - 6 \cdot 3 + 9) = \frac{2}{3\sigma^2},$$

$$E(G_{0i}^0 G_{0i}^2) = \frac{1}{8\sigma^2} (E\tilde{\varepsilon}_i^6 - E\tilde{\varepsilon}_i^4 - 3E\tilde{\varepsilon}_i^2 + 3) =$$

$$= \frac{1}{8\sigma^2} (15 - 3 - 3 + 3) = \frac{3}{2\sigma^2}.$$

$$E((G_{0i}^2)^2) = \frac{1}{16} (E\tilde{\varepsilon}_i^8 - 6E\tilde{\varepsilon}_i^4 + 9) = \frac{1}{16} (105 - 6 \cdot 3 + 9) = 6.$$

Просуммируем по всем наблюдениям и составим блок информационной матрицы, относящийся к  $\mathbf{c}$ . Поскольку информационная матрица блочно-диагональная между  $\mathbf{c}$  и  $\beta$ , то для нахождения интересующей нас статистики достаточно этого блока:

$$\mathcal{I}_{cc} = N \begin{bmatrix} \frac{1}{2\sigma^4} & 0 & \frac{3}{2\sigma^2} \\ 0 & \frac{2}{3\sigma^2} & 0 \\ \frac{3}{2\sigma^2} & 0 & 6 \end{bmatrix}.$$

Обратная матрица:

$$(\mathcal{I}_{cc})^{-1} = \frac{1}{N} \begin{bmatrix} 8\sigma^4 & 0 & -2\sigma^2 \\ 0 & \frac{3}{2}\sigma^2 & 0 \\ -2\sigma^2 & 0 & \frac{2}{3} \end{bmatrix}.$$

Тест множителя Лагранжа равен  $LM = \tilde{\mathbf{g}}_c^\top (\tilde{\mathcal{I}}_{cc})^{-1} \tilde{\mathbf{g}}_c$  и распределен асимптотически как  $\chi^2$  с 2-мя степенями свободы. Градиенты здесь равны ( $\tilde{e}_i$  — нормированные остатки)

$$\tilde{\mathbf{g}}_c = \begin{pmatrix} 0 \\ \frac{1}{3\sigma} \sum_i (3\tilde{e}_i - \tilde{e}_i^3) \\ \frac{1}{4} \sum_i (\tilde{e}_i^4 - 3) \end{pmatrix}.$$

Поэтому

$$LM = \frac{1}{6N} (\sum_i (3\tilde{e}_i - \tilde{e}_i^3))^2 + \frac{1}{24N} (\sum_i (\tilde{e}_i^4 - 3))^2.$$

Два слагаемых, составляющих эту статистику, асимптотически независимы, и каждое распределено как  $\chi^2(1)$ . Первое слагаемое представляет собой тест на асимметрию, а второе — тест на эксцесс. Эту же статистику можно получить и с помощью других семейств распределений. Здесь мы опять сталкиваемся с локально эквивалентными альтернативами.

Точно такой же подход может быть использован в других моделях с нормально распределенными ошибками. Авторы теста Жарк и Бера (Jarque, Bera), применили этот подход к пробиту и моделям с усеченной и цензурированной зависимой переменной.

## Регрессия с ошибками во всех переменных

В классической регрессионной модели ошибка содержится в единственной переменной — той, которая стоит в левой части (зависимой переменной). Рассмотрим более общий случай.

Пусть имеется матрица ненаблюдаемых исходных переменных  $X = (X_1, \dots, X_M)$ ,  $X = \{X_{ij}\}$ ,  $i = 1, \dots, N$ ,  $j = 1, \dots, M$ . Эти переменные связаны между собой линейной зависимостью:  $X\beta = 0$ . Требуется оценить коэффициенты  $\beta$ . Наблюдаются только переменные  $Y$ , которые представляют собой переменные  $X$  измеренные с ошибками:

$$Y_j = X_j + \epsilon_j.$$

Предполагается, что ошибки некоррелированы для разных номеров наблюдений, но ошибки, относящиеся к наблюдениям с одним и тем же номером  $i$  коррелированы, причем матрица ковариаций  $\Omega$  точно известна. Для того, чтобы можно было воспользоваться методом максимального правдоподобия, делаем предположение, что ошибки распределены нормально.

Логарифмическая функция правдоподобия равна

$$\ell = -\frac{N}{2} \ln |\Omega| - \frac{1}{2} \sum_i E_i \Omega^{-1} E_i^T + \text{const} \rightarrow \max_{X, \beta}$$

где  $E_i = Y_j - X_j$  —  $i$ -я строка матрицы остатков.

Максимизируем функцию правдоподобия при ограничении  $X\beta = 0$ .

Задачу максимизации можно записать с помощью лагранжиана:

$$\begin{aligned} L &= -\frac{1}{2} \text{Tr}((Y - X) \Omega^{-1} (Y - X)^T) - \lambda^T X\beta = \\ &= -\frac{1}{2} \text{Tr}(Y \Omega^{-1} Y^T) + \text{Tr}(X \Omega^{-1} Y^T) - \frac{1}{2} \text{Tr}(X \Omega^{-1} X^T) - \lambda^T X\beta. \end{aligned}$$

$$\frac{\partial L}{\partial X} = \Omega^{-1} Y^T - \Omega^{-1} X^T - \beta \lambda^T = 0.$$

Используем

$$\frac{\partial \text{Tr}(AB)}{\partial A} = B \qquad \frac{\partial \text{Tr}(ABA^T)}{\partial A} = 2BA^T \qquad \frac{\partial \text{Tr}(x^T B y)}{\partial A} = y x^T.$$

Отсюда получим выражение для  $X$ :

$$X = Y - \lambda \beta^T \Omega.$$

Если домножить это уравнение на  $\beta$  и вспомнить, что  $X\beta = 0$ , то

$$Y\beta = \lambda \beta^T \Omega \beta, \quad \Rightarrow \quad \lambda = \frac{1}{\beta^T \Omega \beta} Y\beta.$$

Подставим эти соотношения в функцию правдоподобия и получим концентрированную функцию правдоподобия (“концентрированный лагранжиан”):

$$\begin{aligned} \ell^c &= -\frac{1}{2} \text{Tr}(\lambda \beta^T \Omega \Omega^{-1} \Omega \beta \lambda^T) = -\frac{1}{2} \text{Tr}(\lambda^T \lambda \beta^T \Omega \beta) = -\frac{1}{2} \lambda^T \lambda \beta^T \Omega \beta = \\ &= -\frac{1}{2} \frac{\beta^T Y^T Y \beta}{\beta^T \Omega \beta} \end{aligned}$$

Таким образом, нахождение максимума правдоподобия равносильно минимизации следующей функции:

$$\varphi = \frac{\beta^T Y^T Y \beta}{\beta^T \Omega \beta} \rightarrow \min_{\beta}$$

Условие первого порядка для минимума:

$$\begin{aligned} \frac{\partial \varphi}{\partial \beta^T} &= 2 \frac{1}{\beta^T \Omega \beta} Y^T Y \beta - 2 \frac{\beta^T Y^T Y \beta}{(\beta^T \Omega \beta)^2} \Omega \beta = 0 \\ \Rightarrow (Y^T Y - \frac{\beta^T Y^T Y \beta}{\beta^T \Omega \beta} \Omega) \beta &= 0 \end{aligned}$$

или 
$$(\Omega^{-1} Y^T Y - \frac{\beta^T Y^T Y \beta}{\beta^T \Omega \beta}) \beta = (\Omega^{-1} Y^T Y - \varphi) \beta = 0.$$

Таким образом,  $\varphi$  должно быть собственным числом матрицы  $\Omega^{-1} Y^T Y$ , а  $\beta$  — ее собственным вектором. Проверим, что эти два условия не противоречат друг другу. Пусть  $\beta_k$  — некоторый собственный вектор,  $\varphi_k$  — соответствующее собственное число этой матрицы:

$$(\Omega^{-1} Y^T Y - \varphi_k) \beta_k = 0.$$

Покажем, что  $\varphi_k = \frac{\beta_k^T Y^T Y \beta_k}{\beta_k^T \Omega \beta_k}$ .

Домножим слева на  $\beta_k^T \Omega$ :

$$\beta_k^T Y^T Y \beta_k - \varphi_k \beta_k^T \Omega \beta_k = 0.$$

Отсюда получаем требуемое равенство.

Поскольку требуется минимизировать  $\varphi$ , то нужно выбрать минимальное собственное число  $\varphi_{\min}$ . Оценкой вектора коэффициентов  $\hat{\beta}$  будет соответствующий собственный вектор  $\beta_{\min}$ . Отсюда получаем оценки для матрицы исходных переменных  $X$ :

$$X = Y - \lambda \hat{\beta}^T \Omega = Y - \frac{1}{\hat{\beta}^T \Omega \hat{\beta}} Y \hat{\beta} \hat{\beta}^T \Omega = Y \left( I - \frac{1}{\hat{\beta}^T \Omega \hat{\beta}} \hat{\beta} \hat{\beta}^T \Omega \right).$$

В частном случае, когда ошибка имеется только в первой переменной

$$\Omega = \begin{bmatrix} \sigma^2 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{O} \end{bmatrix}.$$

Нужно минимизировать величину

$$\varphi = \frac{\beta^T Y^T Y \beta}{\sigma^2 (\beta_1)^2} = \left( Y_1 + \frac{1}{\sigma \beta_1} Y_{(1)} \beta_{(1)} \right)^T \left( Y_1 + \frac{1}{\sigma \beta_1} Y_{(1)} \beta_{(1)} \right).$$

где  $Y_{(1)}$  — матрица  $Y$  без первого столбца,  $\beta_{(1)}$  — вектор  $\beta$  без первого элемента. Если обозначить  $Y = Y_1$ ,  $X = Y_{(1)}$ ,  $\beta = \beta_{(1)}$  то получим ОМНК:

$$(Y - X\beta)^T (Y - X\beta) \rightarrow \min$$

## **Внешне не связанные регрессионные уравнения**

Пусть есть  $k=1, \dots, K$  регрессионных уравнений и соответствующие выборки одинаковой длины  $i=1, \dots, N$ :

$$Y_k = X_k \beta_k + \epsilon_k, \quad k = 1, \dots, K.$$

Предполагается, что коэффициенты  $\beta_k$  в разных уравнениях не связаны между собой какими-либо ограничениями, т.е. уравнения независимы с точки зрения коэффициентов, но существует связь с точки зрения ошибок наблюдений с одним и тем же номером  $i$ . Т.е. соответствующая корреляционная матрица не является диагональной, причем все элементы этой матрицы неизвестны. Ошибки, относящиеся к наблюдениям с разными номерами, считаются независимыми.

В экономике такое может быть если наблюдения относятся к одному и тому же периоду времени. В этом случае “одновременные” ошибки могут быть сильно коррелированы (как прибыль предприятий из-за экономического цикла). Еще один пример — регрессии, относящиеся к мужьям и женам.

Если обозначить

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_K \end{pmatrix} \quad X = \begin{bmatrix} X_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & X_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \cdots & \cdots & X_K \end{bmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_K \end{pmatrix} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_K \end{pmatrix},$$

тогда модель можно переписать в более компактном виде:

$$Y = X\beta + \epsilon.$$

Пусть остатки  $k$ -й регрессии равны  $e_k(\beta_k) = Y_k - X_k \beta_k$ .

Составим матрицу  $E(\beta)$ , столбцами которой являются  $e_k$ :

$$E_{ki} = Y_{ki} - X_{ki} \beta_{ki}.$$

Строки матрицы  $E$  будем обозначать  $E_i$ .

Предположение модели об ошибках состоит в том, что “одновременные” ошибки коррелированы, а “разновременные” — нет. Таким образом,

$$E(\epsilon_{ki} \epsilon_{li}) = \omega_{kl} \quad \text{и} \quad E(\epsilon_{ki} \epsilon_{ls}) = 0 \quad (i \neq s)$$

или, в матричной записи,

$$E(E_i(\beta_0) E_i(\beta_0)^T) = \Omega \quad \text{и} \quad E(E_i(\beta_0) E_s(\beta_0)^T) = 0 \quad (i \neq s).$$

Составим из ошибок вектор-столбец по другому:

$$\mathbf{\varepsilon}^{\circ} = \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{K1} \\ \vdots \\ \varepsilon_{1N} \\ \vdots \\ \varepsilon_{KN} \end{pmatrix} = \begin{pmatrix} \mathbf{E}_1^{\top}(\boldsymbol{\beta}_0) \\ \vdots \\ \mathbf{E}_N^{\top}(\boldsymbol{\beta}_0) \end{pmatrix},$$

где ошибки, относящиеся к одному и тому же наблюдению  $i$  стоят рядом.

Ковариационная матрица этого вектора ошибок является блочно-диагональной и равна

$$\mathring{V} = E(\mathbf{\varepsilon}^{\circ \top} \mathbf{\varepsilon}^{\circ}) = \begin{bmatrix} \mathbf{\Omega} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Omega} & & \vdots \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \dots & \dots & \mathbf{\Omega} \end{bmatrix}.$$

Обратная к ней тоже блочно-диагональная и равна

$$\mathring{V}^{-1} = \begin{bmatrix} \mathbf{\Omega}^{-1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Omega}^{-1} & & \vdots \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \dots & \dots & \mathbf{\Omega}^{-1} \end{bmatrix}.$$

Если использовать символ Кронекера, то можно записать ковариационную матрицу не переставляя наблюдения:

$$V = E(\boldsymbol{\varepsilon}^{\top} \boldsymbol{\varepsilon}) = \mathbf{\Omega} \otimes \mathbf{I}_N$$

(в этих обозначениях  $\mathring{V} = \mathbf{I}_N \otimes \mathbf{\Omega}$ ), где  $\mathbf{I}_N$  — единичная матрица  $N \times N$ . Аналогично

$$V^{-1} = \mathbf{\Omega}^{-1} \otimes \mathbf{I}_N.$$

Чтобы оценить модель, мы должны указать распределение ошибок. Предполагаем, как обычно, что ошибки имеют нормальное распределение.

$\boldsymbol{\beta}$  и  $\mathbf{\Omega}$  — неизвестные параметры модели.

Так же как и из ошибок, составим один вектор-столбец и из остатков:

$$\mathbf{e}(\boldsymbol{\beta}) = \begin{pmatrix} e_1(\boldsymbol{\beta}) \\ \vdots \\ e_K(\boldsymbol{\beta}) \end{pmatrix}, \quad \mathring{e}(\boldsymbol{\beta}) = \begin{pmatrix} \mathbf{E}_1^{\top}(\boldsymbol{\beta}) \\ \vdots \\ \mathbf{E}_N^{\top}(\boldsymbol{\beta}) \end{pmatrix}$$

Логарифмическая функция правдоподобия равна

$$\begin{aligned}\ell &= -\frac{1}{2} \ln |\mathbf{V}| - \frac{1}{2} \mathbf{e}^\top \mathbf{V}^{-1} \mathbf{e} + \text{const} = \\ &= -\frac{1}{2} \ln |\mathring{\mathbf{V}}| - \frac{1}{2} \mathring{\mathbf{e}}^\top \mathring{\mathbf{V}}^{-1} \mathring{\mathbf{e}} + \text{const} \rightarrow \max_{\boldsymbol{\beta}, \boldsymbol{\Omega}} \\ \ell &= -\frac{N}{2} \ln |\boldsymbol{\Omega}| - \frac{1}{2} \sum_{i=1}^N \mathbf{E}_i \boldsymbol{\Omega}^{-1} \mathbf{E}_i^\top + \text{const} = \\ &= -\frac{N}{2} \ln |\boldsymbol{\Omega}| - \frac{1}{2} \text{Tr}(\mathbf{E} \boldsymbol{\Omega}^{-1} \mathbf{E}^\top) + \text{const}.\end{aligned}$$

В максимуме правдоподобия производная по  $\boldsymbol{\Omega}^{-1}$  равна нулю:

$$\frac{\partial \ell}{\partial \boldsymbol{\Omega}^{-1}} = \frac{N}{2} \boldsymbol{\Omega} - \frac{1}{2} \mathbf{E}^\top \mathbf{E} = 0.$$

Откуда получаем  $\boldsymbol{\Omega}(\boldsymbol{\beta}) = \frac{1}{N} \mathbf{E}(\boldsymbol{\beta})^\top \mathbf{E}(\boldsymbol{\beta})$ .

При дифференцировании мы использовали правила

$$\frac{\partial \ln |\mathbf{A}|}{\partial \mathbf{A}^{-1}} = \mathbf{A}^\top \quad \text{и} \quad \frac{\partial \text{Tr}(\mathbf{A}\mathbf{B}\mathbf{C})}{\partial \mathbf{B}} = \mathbf{C}\mathbf{A}.$$

Подставим  $\boldsymbol{\Omega}(\boldsymbol{\beta})$  в  $\ell(\cdot)$ , чтобы получить концентрированную функцию правдоподобия:

$$\begin{aligned}\ell^c &= -\frac{N}{2} \ln |\boldsymbol{\Omega}(\boldsymbol{\beta})| - \frac{1}{2} \text{Tr}(\mathbf{E} \boldsymbol{\Omega}(\boldsymbol{\beta})^{-1} \mathbf{E}^\top) = -\frac{N}{2} \ln |\boldsymbol{\Omega}(\boldsymbol{\beta})| - \frac{1}{2} \text{Tr}(\mathbf{E}^\top \mathbf{E} \boldsymbol{\Omega}(\boldsymbol{\beta})^{-1}) = \\ &= -\frac{N}{2} \ln |\boldsymbol{\Omega}(\boldsymbol{\beta})| - \text{Tr}(\mathbf{I}_K) = -\frac{N}{2} \ln |\boldsymbol{\Omega}(\boldsymbol{\beta})| - \frac{NK}{2}.\end{aligned}$$

Задача  $\ell^c \rightarrow \max_{\boldsymbol{\beta}}$  эквивалентна задаче

$$\begin{aligned}|\boldsymbol{\Omega}(\boldsymbol{\beta})| &\rightarrow \min_{\boldsymbol{\beta}} \quad \text{или} \\ |\mathbf{E}(\boldsymbol{\beta})^\top \mathbf{E}(\boldsymbol{\beta})| &\rightarrow \min_{\boldsymbol{\beta}}.\end{aligned}$$

Выражение  $\frac{1}{N} |\mathbf{E}(\boldsymbol{\beta})^\top \mathbf{E}(\boldsymbol{\beta})|$  получило название **обобщенной дисперсии**.

Найдем условия первого порядка максимума концентрированной функции правдоподобия.

$$\frac{\partial}{\partial \beta_{kj}} \ln |\boldsymbol{\Omega}(\boldsymbol{\beta})| = \text{Tr}(\boldsymbol{\Omega}(\boldsymbol{\beta})^{-1} \frac{\partial}{\partial \beta_{kj}} \boldsymbol{\Omega}(\boldsymbol{\beta})).$$

Здесь мы используем, что

$$\frac{ds(\mathbf{A})}{dt} = \text{Tr}\left(\frac{\partial s}{\partial \mathbf{A}} \frac{d\mathbf{A}}{dt}\right),$$

где  $s(\mathbf{A})$  — скалярная функция от матрицы  $\mathbf{A}$ , а также что

$$\frac{\partial \ln|\mathbf{A}|}{\partial \mathbf{A}} = \mathbf{A}^{-1}.$$

$$\begin{aligned} \frac{\partial}{\partial \beta_{kj}} \mathbf{\Omega}(\boldsymbol{\beta}) &= \frac{\partial}{\partial \beta_{kj}} \left( \frac{1}{N} \mathbf{E}^\top \mathbf{E} \right) = \frac{2}{N} \mathbf{E}^\top \frac{\partial \mathbf{E}}{\partial \beta_{kj}} = -\frac{2}{N} \mathbf{E}^\top (0, \dots, 0, \mathbf{X}_{kj}, 0, \dots, 0) = \\ &= -\frac{2}{N} (0, \dots, 0, \mathbf{E}^\top \mathbf{X}_{kj}, 0, \dots, 0). \end{aligned}$$

Поэтому,

$$\text{Tr} \left( \mathbf{\Omega}(\boldsymbol{\beta})^{-1} \frac{\partial}{\partial \beta_{kj}} \mathbf{\Omega}(\boldsymbol{\beta}) \right) = -\frac{2}{N} \mathbf{\Omega}(\boldsymbol{\beta})^{-1} \mathbf{E}^\top \mathbf{X}_{kj},$$

где  $\mathbf{\Omega}(\boldsymbol{\beta})^{-1}_k$  —  $k$ -я строка матрицы  $\mathbf{\Omega}(\boldsymbol{\beta})^{-1}$ .

Отсюда

$$\frac{\partial}{\partial \beta_k} \ln|\mathbf{\Omega}(\boldsymbol{\beta})| = -\frac{2}{N} \mathbf{\Omega}(\boldsymbol{\beta})^{-1} \mathbf{E}^\top \mathbf{X}_k = -2 (\mathbf{E}^\top \mathbf{E})^{-1} \mathbf{E}^\top \mathbf{X}_k.$$

Получим условия первого порядка:

$$(\mathbf{E}(\boldsymbol{\beta})^\top \mathbf{E}(\boldsymbol{\beta}))^{-1} \mathbf{E}(\boldsymbol{\beta})^\top \mathbf{X}_k = 0.$$

Эта система уравнений нелинейна относительно  $\boldsymbol{\beta}$ . Один из возможных способов решения состоит в использовании последовательности вспомогательных регрессий. Он основан на том, что эти уравнения для оценок МП  $\hat{\boldsymbol{\beta}}$  (хотя показать это технически сложно) эквивалентны уравнениям

$$\hat{\boldsymbol{\beta}}_k^* = (\mathbf{X}_k^{*\top} \mathbf{X}_k^*)^{-1} \mathbf{X}_k^{*\top} \mathbf{Y}_k,$$

где  $\mathbf{X}_k^* = (\mathbf{X}_k, \mathbf{E}(\hat{\boldsymbol{\beta}})_{-k})$ ,  $\mathbf{E}(\hat{\boldsymbol{\beta}})_{-k}$  — матрица остатков всех уравнений, кроме  $k$ -го.

То есть в каждую регрессию надо добавить остатки из других регрессий.

Оценки вычисляются итеративно, начиная, например, с оценок ОМНК. Стандартные ошибки полученных в результате итераций уравнений надо скорректировать: вычислять не на основе суммы квадратов остатков из вспомогательной регрессии, а на основе суммы квадратов исходных остатков, т. е.  $\mathbf{e}_k(\hat{\boldsymbol{\beta}})^\top \mathbf{e}_k(\hat{\boldsymbol{\beta}}) = (\mathbf{Y}_k - \mathbf{X}_k \hat{\boldsymbol{\beta}})^\top (\mathbf{Y}_k - \mathbf{X}_k \hat{\boldsymbol{\beta}})$ .

В частном случае, когда регрессоры во всех уравнениях одни и те же, оценки МП для коэффициентов  $\boldsymbol{\beta}$  совпадут с оценками ОМНК. Различие будет только в оценке ковариационной матрицы ошибок  $\mathbf{\Omega}$ . Если в условии первого порядка

$$(E(\beta)^T E(\beta))_k^{-1} E(\beta)^T X_k = 0$$

взять  $X_k = X_0 \forall k$ , то должно выполняться  $(E(\beta)^T E(\beta))^{-1} E(\beta)^T X_0 = 0$ , откуда следует, что  $E(\beta)^T X_0 = 0$ , то есть  $e_k(\beta)^T X_0 = 0$  — в каждом уравнении остатки ортогональны матрице регрессоров  $X$ . А это и есть условие минимума суммы квадратов в соответствующем уравнении.

Как и в любой модели обобщенного метода наименьших квадратов информационная матрица параметров  $\beta$  вычисляется по формуле

$$\mathcal{I}_{\beta\beta}^N = X^T V^{-1} X,$$

т.е.

$$\mathcal{I}_{\beta\beta}^N = X^T (\Omega^{-1} \otimes \mathbf{I}_K) X.$$

Другой подход к вычислению оценок МП состоит в использовании итеративного обобщенного МНК.

$$\Omega^t = \frac{1}{N} E(\beta^t)^T E(\beta^t).$$

$$\beta^{t+1} = (X^T ((\Omega^t)^{-1} \otimes \mathbf{I}_K) X)^{-1} X^T ((\Omega^t)^{-1} \otimes \mathbf{I}_K) Y.$$

При использовании этого метода приходится иметь дело с матрицами большой размерности.

## Системы одновременных уравнений

Пусть  $Y$  — матрица  $(N \times m)$  эндогенных переменных,  $X$  — матрица  $(N \times k)$  экзогенных переменных. Предполагается, что они связаны между собой  $m$  уравнениями:

$$Y\Gamma = XB + E.$$

Здесь  $\Gamma$  — квадратная невырожденная матрица  $(m \times m)$  и  $B$   $(k \times m)$  — матрицы коэффициентов.

Такое представление системы одновременных уравнений называют **структурной формой**. Коэффициенты здесь можно определить только с точностью до множителей. Если каждое уравнение пронормировать, то изменятся только неизвестные параметры в  $\Gamma$ ,  $B$  и матрице ковариаций ошибок. Один из способов нормировки заключается в том, чтобы взять диагональные элементы матрицы  $\Gamma$  равными 1:

$$\Gamma_{ll} = 1, \quad l = 1, \dots, m.$$

Предполагается кроме того, что часть коэффициентов в матрицах  $\Gamma$  и  $B$  могут быть равны нулю. Если оставить в уравнениях только неизвестные ненулевые коэффициенты, то можно переписать их в виде

$$Y_l = Y_{-l} \gamma_l + X_l \beta_l + \epsilon_l, \quad l = 1, \dots, m.$$

В каждом уравнении в левой части остается только одна эндогенная переменная, имеющая тот же номер, что и уравнение. Остальные эндогенные переменные с ненулевыми коэффициентами перенесены в правую часть.  $Y_{-l}$  — составленная из них матрица.  $X_l$  — это экзогенные переменные с ненулевыми коэффициентами.

Систему одновременных уравнений можно записать также в приведенной форме, где каждая эндогенная переменная представлена как функция только экзогенных переменных:

$$Y = X\Pi + U.$$

Структурной форме соответствует ограниченная приведенная форма, называемая так потому, что на коэффициенты накладываются ограничения  $\Pi = B\Gamma^{-1}$ , так что:

$$Y = X\Pi + U = XB\Gamma^{-1} + E\Gamma^{-1}.$$

Если ограничения не учитываются, то имеем **неограниченную приведенную форму**. Неограниченная приведенная форма представляет собой систему внешне не связанных уравнений с одной и той же матрицей регрессоров во

всех уравнениях. Таким образом, коэффициенты в ней можно оценить с помощью ОМНК.

Рассмотрим два подхода к оцениванию системы одновременных уравнений методом максимального правдоподобия.

### FIML

Метод максимального правдоподобия, использующий полную информацию, (full information maximum likelihood) называется так, потому что он использует всю информацию об ограничениях, в том числе информацию о том, что часть коэффициентов равна нулю и  $\Pi = B \Gamma^{-1}$ .

Для применения ММП предположим, что  $E_i \sim \text{NID}(\mathbf{0}, \Omega)$ , то есть ошибки имеют нормальное распределение; ошибки, относящиеся к одному и тому же номеру наблюдения  $i$  коррелированы с матрицей ковариаций  $\Omega$ , а относящиеся к наблюдениям с разными номерами — некоррелированы.

Матрицы  $\Gamma$ ,  $B$ ,  $\Omega$  нужно оценить. Функция правдоподобия для  $i$ -го наблюдения равна

$$\ell_i = -\frac{m}{2} \ln 2\pi - \frac{1}{2} \ln |\Omega| - \frac{1}{2} E_i \Omega^{-1} E_i^\top.$$

Заменим  $E_i$  на  $Y_i \Gamma - X_i B$ , при этом в функцию правдоподобия нужно добавить якобианный член, соответствующий преобразованию  $Y_i$  в  $E_i$ . Якобиан этого преобразования равен

$$J = \frac{dY_i}{dE_i} = \Gamma.$$

Таким образом, якобианный член равен

$$\ln (\text{abs} | \Gamma |).$$

Просуммируем функции правдоподобия отдельных наблюдений:

$$\begin{aligned} \ell = \sum_i \ell_i = & -\frac{Nm}{2} \ln 2\pi + N \ln (\text{abs} | \Gamma |) - \frac{N}{2} \ln |\Omega| - \\ & - \frac{1}{2} \sum_i E_i(\Gamma, B) \Omega^{-1} E_i(\Gamma, B)^\top. \end{aligned}$$

Концентрируем функцию правдоподобия по  $\Omega$ . В максимуме

$$\frac{\partial \ell}{\partial \Omega^{-1}} = \frac{N}{2} \Omega - \frac{1}{2} \sum_i E_i^\top E_i = \frac{N}{2} \Omega - \frac{1}{2} E^\top E = 0.$$

Отсюда имеем:

$$\Omega(\Gamma, \mathbf{B}) = \frac{1}{N} (\mathbf{Y}\Gamma - \mathbf{X}\mathbf{B})^\top (\mathbf{Y}\Gamma - \mathbf{X}\mathbf{B}).$$

Концентрированная функция правдоподобия равна

$$\ell^c = N \ln (\text{abs}|\Gamma|) - \frac{N}{2} \ln |(\mathbf{Y}\Gamma - \mathbf{X}\mathbf{B})^\top (\mathbf{Y}\Gamma - \mathbf{X}\mathbf{B})| + \text{const.}$$

Можно переписать ее в другом виде:

$$\ell^c = -\frac{N}{2} \ln \left| \frac{1}{|\Gamma|^2} (\mathbf{Y}\Gamma - \mathbf{X}\mathbf{B})^\top (\mathbf{Y}\Gamma - \mathbf{X}\mathbf{B}) \right| + \text{const} =$$

$$= -\frac{N}{2} \ln |(\mathbf{Y} - \mathbf{X}\mathbf{B}\Gamma^{-1})^\top (\mathbf{Y} - \mathbf{X}\mathbf{B}\Gamma^{-1})| + \text{const.}$$

Последнее выражение совпадает с концентрированной по матрице ковариаций функцией правдоподобия для ограниченной приведенной формы. Причина этого заключается в том, что структурная и ограниченная приведенная формы являются только различными способами записи одной и той же модели.

Методы оценивания систем одновременных уравнений довольно громоздки. Хорошие результаты дает применение метода Ньютона, но выражение для гессиана имеет довольно сложный вид.

Опишем здесь один из возможных методов, который имеет интуитивно понятную интерпретацию, и который несложно реализовать в виде компьютерной программы.

“Очищенные” от ошибок переменные  $\mathbf{Y}$  можно определить как

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{B}\Gamma^{-1}.$$

Матрица соответствующих остатков

$$\hat{\mathbf{U}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\mathbf{B}\Gamma^{-1}.$$

Используя эти обозначения, уравнения системы

$$\mathbf{Y}_l = \mathbf{Y}_{-l} \gamma_l + \mathbf{X}_l \beta_l + \varepsilon_l$$

можно переписать в виде

$$\mathbf{Y}_l - \hat{\mathbf{U}}_{-l} \gamma_l = \hat{\mathbf{Y}}_{-l} \gamma_l + \mathbf{X}_l \beta_l + \varepsilon_l,$$

где  $\hat{\mathbf{Y}}_{-l}$  — это “очищенные” переменные  $\mathbf{Y}_{-l}$ , и  $\hat{\mathbf{U}}_{-l} = \mathbf{Y}_{-l} - \hat{\mathbf{Y}}_{-l}$ . Если рассматривать в этих уравнениях  $\hat{\mathbf{U}}_{-l} \gamma_l$  и  $\hat{\mathbf{Y}}_{-l}$  как известные, то получаем систему внешне не связанных регрессионных уравнений. Все случайные компоненты как бы переносятся в левую часть регрессионных уравнений, так как если

переменные  $\hat{Y}_{-l}$  вычисляются на основании состоятельных оценок параметров  $\mathbf{\Gamma}$  и  $\mathbf{B}$ , то они асимптотически некоррелированы с ошибками.

К этим уравнениям можно применить один из итеративных методов оценивания внешне не связанных регрессий. Величины  $\hat{U}_{-l}$  и  $\hat{Y}_{-l}$  вычисляются на основании оценок параметров  $\mathbf{\Gamma}$  и  $\mathbf{B}$ , полученных на предыдущих итерациях. Как можно показать, этот алгоритм сходится к оценкам максимального правдоподобия.

После того, как получены оценки FIML, интересно сравнить структурную форму с неограниченной приведенной формой:  $\mathbf{Y} = \mathbf{X}\mathbf{\Pi} + \mathbf{U}$ .

Как уже говорилось, неограниченную форму можно оценить, применяя ОМНК к каждому уравнению, т. е.,  $\hat{\mathbf{\Pi}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ , только в качестве оценки ковариационной матрицы ошибок следует взять

$$\begin{aligned} \hat{\mathbf{\Omega}} &= \frac{1}{N} \mathbf{U}(\hat{\mathbf{\Pi}})^T \mathbf{U}(\hat{\mathbf{\Pi}}) = \frac{1}{N} (\mathbf{Y} - \mathbf{X}\hat{\mathbf{\Pi}})^T (\mathbf{Y} - \mathbf{X}\hat{\mathbf{\Pi}}) \\ &= \frac{1}{N} \mathbf{Y}^T (\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) \mathbf{Y} = \frac{1}{N} \mathbf{Y}^T \mathbf{M}_X \mathbf{Y}. \end{aligned}$$

Логарифмическая функция правдоподобия в максимуме для ограниченной и неограниченной модели равна соответственно

$$\begin{aligned} \hat{\ell} &= -\frac{N}{2} \ln |(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}\hat{\mathbf{\Gamma}}^{-1})^T (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}\hat{\mathbf{\Gamma}}^{-1})| + \text{const} \\ \text{и} \quad \tilde{\ell} &= -\frac{N}{2} \ln |\mathbf{Y}^T \mathbf{M}_X \mathbf{Y}| + \text{const}. \end{aligned}$$

Константа в обеих формулах одна и та же и равна  $-\frac{Nm}{2} \ln 2\pi + \frac{N}{2} \ln N$ .

Статистика отношения правдоподобия, равная  $LR = 2(\tilde{\ell} - \hat{\ell})$  имеет распределение  $\chi^2$  с числом степеней свободы равным количеству ограничений. Этот тест называется **тестом на сверхидентифицирующие ограничения**.

Поскольку неограниченную форму оценивать легче, то имеет смысл использовать ее для проверки различных гипотез. Если ограничения выполнены, то оценивая неограниченную модель мы теряем в эффективности, но оценки все же будут состоятельными. Удобно проверять таким образом регрессию на автокорреляцию остатков, гетероскедастичность, функциональную форму.

## LIML

Метод максимального правдоподобия, использующий ограниченную информацию, (limited information maximum likelihood) предназначен для оценивания одного уравнения из системы одновременных уравнений. Остальные уравнения оцениваются только в той степени, в какой это требуется для оценивания первого уравнения. Первое уравнение оценивается в структурной форме, а остальные — в неограниченной приведенной (тем самым, используется не вся имеющаяся информация):

$$\begin{aligned} Y_1 - Y_{-1} \gamma_1 &= X_1 \beta_1 + \varepsilon_1, \\ Y_{-1} &= X_1 B_1 + X_{-1} B_{-1} + E_{-1}. \end{aligned}$$

Здесь  $Y_1$  — “зависимая” переменная в первом уравнении,  $Y_{-1}$  — другие эндогенные переменные, входящие в первое уравнение,  $X_1$  — экзогенные переменные, входящие в первое уравнение,  $X_{-1}$  — остальные экзогенные переменные системы.

Удобно рассматривать данную систему уравнений как структурную форму с ограничениями на матрица  $\Gamma$  и  $B$ .

Обозначим

$$\begin{aligned} Y &= (Y_1, Y_{-1}), X = (X_1, X_{-1}), \gamma = \begin{bmatrix} 1 \\ -\gamma_1 \end{bmatrix}, \\ M_X &= I - X(X^T X)^{-1} X^T, \quad M_1 = I - X_1(X_1^T X_1)^{-1} X_1^T. \end{aligned}$$

Задача нахождения оценок максимального правдоподобия сводится к задаче нахождения **наименьшего дисперсионного отношения**:

$$\frac{\gamma^T Y^T M_X Y \gamma}{\gamma^T Y^T M_1 Y \gamma} \rightarrow \min_{\gamma}$$

Условие первого порядка минимума:

$$\begin{aligned} 2 \frac{1}{\gamma^T Y^T M_1 Y \gamma} Y^T M_1 Y \gamma - 2 \frac{\gamma^T Y^T M_X Y \gamma}{(\gamma^T Y^T M_1 Y \gamma)^2} Y^T M_X Y \gamma &= 0. \\ \Rightarrow (Y^T M_1 Y - \frac{\gamma^T Y^T M_X Y \gamma}{\gamma^T Y^T M_1 Y \gamma} Y^T M_X Y) \gamma &= 0. \end{aligned}$$

Отсюда следует, что  $\phi = \frac{\gamma^T Y^T M_X Y \gamma}{\gamma^T Y^T M_1 Y \gamma}$  — собственное число матрицы  $Y^T M_1 Y (Y^T M_X Y)^{-1}$ , а  $Y^T M_X Y \gamma$  — соответствующий собственный вектор. Поскольку удобно искать собственные числа симметричной матрицы, то лучше взять матрицу

$$(Y^T M_X Y)^{-1/2} Y^T M_1 Y (Y^T M_X Y)^{-1/2},$$

которая имеет те же собственные числа.

Таким образом, метод наименьшего дисперсионного отношения сводится к задаче отыскания минимального собственного числа вещественной симметричной матрицы.

## **Использованная литература**

- Дрейпер, Н., Смит, Г. *Прикладной регрессионный анализ*: В 2-х кн. — М: Финансы и статистика, 1986.
- Песаран, М., Слейтер, Л. *Динамическая регрессия: теория и алгоритмы*. — М: Финансы и статистика, 1984.
- *Статистические методы в экспериментальной физике*. — М: Атомиздат, 1976.
- Amemiya, T. "Selection of Regressors," *International Economic Review*, **21** (1980), 331-354.
- Beggs, J.J. "Diagnostic Testing in Applied Econometrics," *Economic Record*, **64** (1988), 81-101.
- Bera A.K., C.M. Jarque, and L.-F. Lee. "Testing the Normality Assumption in Limited Dependent Variable Models," *International Economic review*, **25** (1984), 563-578.
- Bollerslev, T. "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics*, **31** (1986), 307-327.
- Cramer, J.S. *The Logit Model for Economists*. Adward Arnold, 1991.
- Dagenais, M.G. "The Computation of FIML Estimates as Iterative Generalized Least Squares Estimates in Linear and Nonlinear Simultaneous Equations Models," *Econometrica*, **46** (1978), 1351-1362.
- Davidson, R., and J.G. MacKinnon. "Convenient Specification Tests for Logit and Probit Models," *Journal of Econometrics*, **25** (1984), 241-262.
- Davidson, R., and J.G. MacKinnon. *Estimation and Inference in Econometrics*. Oxford University Press, 1993.
- Enders, W. *Applied Econometric Time Series*. John Wiley & Sons inc., 1995.
- Engle, R.F. "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, **50** (1982), 987-1007.
- Engle, R.F. "Band Spectrum Regression," *Economic Review*, **15** (1974), 1-11.
- Fiorentini, G., G. Calzolari, and L. Panattoni. "Analitical Derivatives and the Computation of GARCH Estimates," *Journal of Applied Econometrics*, **11** (1996), 399-417.
- Godfrey, L.G. *Misspecification Tests in Econometrics: The Lagrange Multiplier Principle and Other Approaches*. Cambridge University Press, 1988.
- Granger, C.W.J., and P.Newbold. "Spurious Regressions in Econometrics," *Journal of Econometrics*, **21** (1974), 111-120.

- Jarque, C.M, and A.K. Bera. “Efficient Tests for Normality, Homoskedasticity, and Serial Independence of Regression Residuals,” *Economic Letters*, 6 (1980), 255-259.
- Krämer, W., and H. Sonnberger. *The Linear Regression Model Under Test*. Physica-Verlag, 1986.
- MacKinnon, J.G., A.A. Haug, and L. Michelis. “Numerical Distribution Functions of Likelihood Ratio Tests for Cointegration” (discussion paper), 1996.
- Pagan, A.R., and D.F. Nicholls. “Estimating Predictions, Prediction Errors and Their Standard Deviations Using Costructed Variables,” *Journal of Econometrics*, **24** (1984), 293-310.
- Pesaran, M.H., and B. Pesaran. *Microfit 3.0. An Interactive Econometric Software Package (User Manual)*. Oxford University Press, 1991.
- Telser L.G. “Iterative Estimation of a Set of Linear Regression Equasions,” *Journal of American Statistical Asociacion*, **59**, 845-862.

## **Предметный указатель**

**A**

**ARCH • 53**

**B**

**ВННН • 18, 20**

**G**

**GARCH • 54**

**M**

**method of scoring • 20**

**A**

**асимптотическая информационная  
матрица • 5**

**асимптотическая эквивалентность трех  
классических тестов • 25**

**асимптотические t- и F-статистики • 29**

**B**

**Бокса-Кокса модель • 61**

**Бокса-Кокса преобразование • 60**

**B**

**взвешенная регрессия • 39**

**вкладов в градиент матрица • 13**

**вклады отдельных наблюдений в  
функцию правдоподобия • 4**

**внешне не связанные регрессионные  
уравнения • 70**

**внешнее произведение градиента • 18**

**вспомогательная регрессия • 11**

для модели Бокса-Кокса • 61

**G**

**Гаусса-Ньютона регрессия**

определение • 45

**гессиан логарифмической функции  
правдоподобия • 4**

эмпирический • 18

**гетероскедастичность • 39**

**градиент логарифмической функции  
правдоподобия • 4**

**градиентные методы • 20**

**I**

**инвариантность ММП • 9**

**информационная матрица**

вычисление • 14

определение • 5

способы оценивания • 17

**искусственная регрессия • 11**

для логита • 31

для пробита • 32

для пуассоновой регрессии • 34

для регрессии с мультипликативной  
гетероскедастичностью • 44

метод ВННН (OPG) • 20

метод Гаусса-Ньютона • 45

**истинное распределение • 3**

**истинный параметр • 3**

**K**

**квази-МП методы • 11, 18**

**ковариационная матрица оценок МП • 17**

**концентрированная функция  
правдоподобия • 21**

**Кронекера произведение • 71**

**L**

**логарифмическая функция правдоподобия  
• 3**

**локально эквивалентные альтернативы •  
42**

**M**

**максимального правдоподобия метод • 3**

**максимального правдоподобия оценки**

альтернативное определение • 5

асимптотическое распределение • 17

вычисление • 20

определение • 3

**мультипликативная**

гетероскедастичность • 42

**Н**  
**наименьшего дисперсионного отношения**  
**метод • 79**

**Ньютона метод • 20**

**О**  
**обобщенная дисперсия • 72**

**П**  
**пирсоновское семейство распределений •**  
**63**

**порождающий данные процесс • 3**

**приведенная форма системы**  
**одновременных уравнений • 75**

**Прэйса-Винстена преобразование • 48**

**Пуассона распределение • 33**

**Р**

**регрессия**

пуассоновская • 34

с AR-ошибкой • 46

с MA-ошибкой • 50

с мультипликативной  
гетероскедастичностью • 42

**С**

**семейство распределений • 3**

**структурная форма системы**  
**одновременных уравнений • 75**

**Т**

**тест**

Вальда • 25

множителя Лагранжа • 23

множителя Лагранжа в градиентной форме  
• 24

на авторегрессионную условную

гетероскедастичность • 57

на гетероскедастичность • 42

на нормальность • 66

на сверхидентифицирующие ограничения  
• 78

отношения правдоподобия • 24

**У**

**уравнения правдоподобия • 4**

**Ф**

**функция правдоподобия • 3**

**Х**

**Холецкого разложение • 37**

**Я**

**якобиан преобразования плотности**  
**распределения • 59**