

Алгебра и геометрия МНК	2
Геометрическая интерпретация МНК	6
Показатель влиятельности	7
Теорема о разбиении регрессоров	9
Вычисление оценок МНК и других регрессионных величин	13
Ортогональная матрица регрессоров	18
Линейные преобразования переменных регрессии	20
Линейная регрессия как вероятностная модель.	
Свойства оценок	21
Свойства оценок МНК в конечных выборках	25
Асимптотические свойства оценок МНК:	
состоятельность	30
Свойства функций от коэффициентов МНК	37
Асимптотические свойства оценок МНК: сходимость в среднеквадратическом	38
Следствия нормальности ошибок	41
Линейные ограничения и проверка гипотез в регрессии	44
Оценки МНК при линейных ограничениях	44
Проверка статистических гипотез	47
Проверка линейных гипотез в регрессии	52
Критерии удаления переменных	57
Критерии правильности спецификации и критерий добавления переменной	60
Гипотезы, лежащие в основе МНК, и их невыполнение	61
Функциональная форма	63
Оценка с предварительной проверкой гипотезы	67
Нарушение гипотезы об ортогональности	68
Идентифицируемость	69
Сферичность	70
Автокорреляция ошибок (сериальная корреляция)	70
Гетероскедастичность	75
Взвешенная регрессия	78
Обобщенный МНК (метод Эйткена???)	80
Нормальность	81

## Алгебра и геометрия МНК

Модель линейной регрессии:

$$y_i = \beta_1 x_{i1} + \dots + \beta_m x_{im} + \varepsilon_i, \quad i = 1, \dots, n.$$

Здесь  $i$  — номер наблюдения. Предполагаем, что всего имеется  $n$  наблюдений.

Зависимая переменная:

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

Матрица регрессоров:

$$X = \begin{pmatrix} x_{11} & \dots & x_{1m} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{nm} \end{pmatrix}.$$

Когда говорят об экспериментах, то матрицу  $X$  называют матрицей плана.

Коэффициенты регрессии — параметры:

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}.$$

Ошибки:

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Уравнение регрессии в матричной форме:

$$y = X\beta + \varepsilon.$$

То же уравнение бывает удобно записать в векторной форме:

$$y = \beta_1 x_1 + \dots + \beta_m x_m + \varepsilon.$$

Здесь

$$x_j = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix}.$$

Вектора  $x_j$  называют регрессорами, а также факторами, независимыми переменными, объясняющими переменными. Они являются столбцами матрицы  $X$ , почему она и называется матрицей регрессоров:

$$X = (x_1, \dots, x_m).$$

Часто матрицу  $X$  бывает удобно разбить на строки

$$X_i = (x_{i1}, \dots, x_{im}),$$

т.е.

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}.$$

Остаток по  $i$ -му наблюдению, соответствующий вектору коэффициентов  $\beta$ :

$$\varepsilon_i(\beta) = y_i - X_i\beta.$$

(В отличие от ошибок, обозначаемых просто через  $\varepsilon_i$ , будем записывать остатки как функцию от аргумента  $\beta$ , т.е.  $\varepsilon_i(\beta)$ .)

Вектор остатков:  $\varepsilon(\beta) = y - X\beta$

Метод наименьших квадратов (МНК) состоит в минимизации суммы квадратов остатков ( $RSS$  — residual sum of squares) по  $\beta$ :

$$RSS(\beta) = \sum_{i=1}^n \varepsilon_i^2(\beta) = \sum_{i=1}^n (y_i - X_i\beta)^2 \rightarrow \min_{\beta}.$$

Сумма квадратов остатков в матричном виде:

$$\begin{aligned} RSS(\beta) &= \varepsilon(\beta)^T \varepsilon(\beta) = (y - X\beta)^T (y - X\beta) = \\ &= y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta = y^T y - 2y^T X\beta + \beta^T X^T X\beta. \end{aligned}$$

Используя матричное (в данном случае векторное) дифференцирование, можно вычислить вектор первых производных суммы квадратов остатков по коэффициентам регрессии:

$$\frac{dRSS(\beta)}{d\beta} = \left( \frac{dRSS(\beta)}{d\beta_1}, \dots, \frac{dRSS(\beta)}{d\beta_m} \right) = -2y^T X + 2\beta^T X^T X.$$

Пусть, например, минимум достигается при  $\beta = \hat{\beta}$ . Тогда в этой точке должно быть выполнено условие первого порядка:

$$\frac{dRSS(\hat{\beta})}{d\beta} = -2y^T X + 2\hat{\beta}^T X^T X = 0^T.$$

После несложных преобразований это условие первого порядка можно записать в виде:

$$X^T(y - X\hat{\beta}) = 0.$$

или

$$X^T y = X^T X \hat{\beta}.$$

Данное соотношение обычно называют нормальным уравнением.

Условие первого порядка означает, что остатки ортогональны регрессорам:

$$X^T e = 0$$

или

$$x_j^T e = 0, j = 1, \dots, m.$$

Здесь  $e = \varepsilon(\beta) = y - X\beta$  — вектор остатков МНК.

В частности, если один из регрессоров (например 1-й) является вектором, состоящим из единиц (т.е.  $x_1 = \mathbb{1}$ ), или, другими словами, в регрессии имеется константа (или свободный член) то

$$x_1^T e = \mathbb{1}^T e = \sum_{i=1}^n e_i = 0.$$

Т.е. в этом случае сумма остатков равна нулю. Среднее остатков тоже будет равно нулю:

$$\bar{e} = \sum_{i=1}^n e_i = 0.$$

Для того, чтобы убедиться, что нормальные уравнения действительно определяют минимум, нужно убедиться, что матрица вторых производных (матрица Гессе) положительно полуопределена. Матрица Гессе для  $RSS(\beta)$  равна

$$\begin{aligned} \frac{d^2 RSS(\beta)}{d\beta d\beta^T} &= \frac{d^2}{d\beta d\beta^T} (y^T y - 2y^T X\beta + \beta^T X^T X\beta) = \\ &= \frac{d}{d\beta^T} (-2y^T X + 2\beta^T X^T X) = 2X^T X. \end{aligned}$$

Эта матрица не зависит от  $\beta$ , всегда одна и та же. Кроме того, она положительно полуопределена. Если к тому же  $\det(X^T X) \neq 0$ , то положительно определена. Поскольку матрица Гессе всюду положительно (полу-) определена, то условия 1-го порядка определяют глобальный минимум суммы квадратов остатков  $RSS(\beta)$ .

Предположение  $\det(X^T X) \neq 0$ , гарантирует единственность минимума. Условие  $\det(X^T X) \neq 0$  эквивалентно тому, что матрица  $X$  имеет полный ранг по столбцам

$$\text{rank}(X) = m,$$

т.е. оно выполнено тогда и только тогда, когда количество регрессоров не превышает количества наблюдений ( $n \geq m$ ), и регрессоры (столбцы матрицы  $X$ ) линейно независимы. Далее мы, как правило, будем считать, что это предположение выполнено (невыврожденный случай).

Из нормального уравнения в предположении невырожденности можем найти вектор коэффициентов МНК:

$$\begin{aligned} X^T y &= X^T X \hat{\beta} \Rightarrow \\ \hat{\beta} &= (X^T X)^{-1} X^T y. \end{aligned}$$

Остатки МНК равны

$$e = y - X\hat{\beta}.$$

Если  $\det(X^T X) = 0$ , то решение не единственное. На самом деле решений в этом случае бесконечно много (континуум). Это означает, что имеющиеся наблюдения не позволяют получить оценки МНК однозначно. Можно назвать это неидентифицируемостью, поскольку в этом случае даже если бы нам были точно известны значения величин  $X_i\beta$ , мы не могли бы по ним восстановить вектор  $\beta$ .

Хотя при  $\det(X^T X) = 0$  оценки МНК не единственны, но условие первого порядка  $X^T e = 0$  все равно выполнено, и остатки однозначно определяются этим условием. Остатки можно вычислить следующим образом. Объединим максимальный линейно независимый набор регрессоров в матрицу  $X^*$ . Тогда если  $e^*$  — остатки из регрессии  $y$  по  $X^*$  (т.е.  $e^* = y - X^*\hat{\beta}^*$ , где  $\hat{\beta}^* = (X^{*T} X^*)^{-1} X^{*T} y$ ), то  $e^* = e$ .

Точно также всегда однозначно определяются расчетные значения  $X\hat{\beta}$  зависимой переменной  $y$ , которые обозначаются  $\hat{y}$ :

$$\hat{y} = X\hat{\beta} = y - e.$$

В невырожденном случае

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

Расчетные значения  $\hat{y}_i$  можно рассматривать как прогноз величины  $X_i\beta$  (т.е.  $y_i$  за вычетом  $\varepsilon_i$ ), либо как прогноз  $y_i$  в точке  $X_i$ .

Так как  $X^T e = 0$ , то

$$\hat{y}^T e = 0,$$

т.е. остатки и расчетные значения ортогональны.

Поскольку  $y = \hat{y} + e$ , то

$$y^T y = \hat{y}^T \hat{y} + e^T e.$$

Действительно, используя,  $\hat{y}^T e = 0$ ,

$$\begin{aligned} y^T y &= (\hat{y} + e)^T (\hat{y} + e) = \hat{y}^T \hat{y} + \hat{y}^T e + e^T \hat{y} + e^T e = \\ &= \hat{y}^T \hat{y} + e^T e. \end{aligned}$$

Получили разложение суммы квадратов:

$$TSS = ESS + RSS.$$

Общая сумма квадратов (англ. total sum of squares):

$$TSS = \mathbf{y}^T \mathbf{y}.$$

Объясненная сумма квадратов (англ. explained sum of squares):

$$ESS = \hat{\mathbf{y}}^T \hat{\mathbf{y}} = \mathbf{y}^T \mathbf{y} - \mathbf{e}^T \mathbf{e}.$$

Сумма квадратов остатков (остаточная сумма квадратов, англ. residual sum of squares):

$$RSS = \mathbf{e}^T \mathbf{e}.$$

### Геометрическая интерпретация МНК

Рассмотрим евклидово пространство  $E^n$ . (Заметим, что  $\mathbf{y}$ ,  $\mathbf{e}$ ,  $\hat{\mathbf{y}}$ ,  $\mathbf{e}$  и регрессоры являются векторами длины  $n$ ). Обозначим  $\mathcal{L}(X)$  подпространство в  $E^n$ , натянутое на регрессоры  $\mathbf{x}_1, \dots, \mathbf{x}_m$ .

Геометрически задача МНК состоит в том, чтобы найти такой вектор  $\hat{\mathbf{y}}$  из  $\mathcal{L}(X)$ , чтобы евклидово расстояние между  $\mathbf{y}$  и  $\hat{\mathbf{y}}$  было минимальным. Иными словами, мы ищем среди всех линейных комбинаций регрессоров наиболее близкую к  $\mathbf{y}$ :

$$\begin{aligned} \|\mathbf{y} - \hat{\mathbf{y}}\| &\rightarrow \min \\ \hat{\mathbf{y}} &\in \mathcal{L}(X). \end{aligned}$$

Эта задача эквивалентна задаче МНК, поскольку:

(1) Минимизация расстояния эквивалентна минимизации квадрата расстояния  $\|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ .

(2) Условие  $\hat{\mathbf{y}} \in \mathcal{L}(X)$  эквивалентно тому, что существует вектор  $\boldsymbol{\beta} \in \mathbb{R}^m$ , такой что  $\hat{\mathbf{y}} = X\boldsymbol{\beta}$ .

В точке минимума остатки  $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$  ортогональны (перпендикулярны) подпространству  $\mathcal{L}(X)$ , и  $\hat{\mathbf{y}}$  есть проекция  $\mathbf{y}$  на  $\mathcal{L}(X)$ .

Обозначим матрицу проекции на подпространство  $\mathcal{L}(X)$  через  $P_X$  или просто  $P$ . Используя это обозначение, можно записать

$$\hat{\mathbf{y}} = P\mathbf{y}.$$

В невырожденном случае

$$P = X(X^T X)^{-1} X^T.$$

Матрицу  $P$  называют иногда «матрицей крышки», поскольку она «добавляет крышку над  $y$ ».

Можно рассмотреть также подпространство  $\mathcal{L}^\perp(X)$ , являющееся ортогональным дополнением  $\mathcal{L}(X)$  в  $E^n$ . Обозначим  $M_X$  или просто  $M$  матрицу проекции на  $\mathcal{L}^\perp(X)$ .

Остатки удовлетворяют соотношению

$$e = My,$$

т.е. они являются проекцией  $y$  на  $\mathcal{L}^\perp(X)$ .

В невырожденном случае

$$M = I_n - X(X^T X)^{-1} X^T.$$

Свойства матриц  $P$  и  $M$ :

1) Матрицы  $P$  и  $M$  однозначно задаются матрицей  $X$  (даже в вырожденном случае  $\text{rank}(X) < m$ ).

2) Матрицы  $P$  и  $M$  симметричны:

$$P^T = P, \quad M^T = M.$$

3) Матрицы  $P$  и  $M$  идемпотентны:

$$PP = P^2 = P, \quad MM = M^2 = M.$$

4) Матрицы  $P$  и  $M$  «погашают» друг друга:

$$PM = 0.$$

5)  $P + M = I_n$ .

6)  $\text{rank}(P) + \text{rank}(M) = n$ .

7)  $PX = X, \quad MX = 0$ .

Из свойств симметричных идемпотентных матриц следуют следующие два свойства:

8)  $\text{rank}(P) = \text{tr}(P)$  и  $\text{rank}(M) = \text{tr}(M)$  (ранг и след матриц  $P$  и  $M$  равны между собой).

9) Все собственные числа матриц  $P$  и  $M$  — нули либо единицы, причем единиц ровно столько, каков ранг матрицы.

Докажем свойство (8) в невырожденном случае:

$$\begin{aligned} \text{tr}(P) &= \text{tr}(X(X^T X)^{-1} X^T) = \text{tr}(X^T X(X^T X)^{-1}) = \\ &= \text{tr}(I_m) = m = \text{rank}(P) \end{aligned}$$

и

$$\text{tr}(M) = \text{tr}(I_n - P) = \text{tr}(I_n) - \text{tr}(P) = n - m = \text{rank}(M).$$

## Показатель влиятельности

Как изменится  $\hat{y}_i$ , если  $y_i$  изменится на величину  $\Delta y_i$ ?

Вспомним, что  $\hat{\mathbf{y}} = \mathbf{P}\mathbf{y}$ , т.е.

$$\hat{y}_i = \sum_{k=1}^n P_{ik} y_k = \sum_{k \neq i} P_{ik} y_k + P_{ii} y_i.$$

Таким образом,

$$\Delta \hat{y}_i = P_{ii} \Delta y_i.$$

Здесь —  $i$ -й диагональный элемент проекционной матрицы  $\mathbf{P}$ . Обозначим его  $h_i$ .

Число  $h_i$  называют показателем влияния или DFFITS (англ.???).

$$h_i = \frac{\Delta \hat{y}_i}{\Delta y_i}.$$

Показатель влияния всегда положителен (в невырожденном случае), поскольку он является квадратичной формой с положительно определенной матрицей:

$$h_i = \mathbf{X}_i (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}_i^\top > 0.$$

Матрица  $(\mathbf{X}^\top \mathbf{X})^{-1}$  положительно определена, поскольку является обратной к положительно определенной матрице  $\mathbf{X}^\top \mathbf{X}$ .

Заметим, что

$$\sum_{i=1}^n h_i = \text{tr}(\mathbf{P}) = \text{rank}(\mathbf{P}) = m.$$

Средняя величина  $h_i$  равна

$$\bar{h} = \frac{1}{n} \sum_{i=1}^n h_i = \frac{m}{n}.$$

Если  $h_i$  превышает  $\bar{h}$  в несколько раз (например, в 4 раза), то наблюдение  $i$  можно считать влиятельным. Показатель влияния для «сбалансированной» матрицы  $\mathbf{X}$  должен быть таким, что  $h_i \approx \bar{h} \forall i$ .

Можно также посмотреть, как влияет изменение  $y_i$  на коэффициенты  $\hat{\beta}_j$ . Вектора  $\mathbf{y}$  и  $\hat{\beta}$  связаны соотношением

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

Введем обозначение  $\mathbf{C} = \{c_{ij}\} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ . Тогда

$$\hat{\beta}_j = \sum_{i=1}^n c_{ij} y_i.$$

Если  $y_i$  изменится на величину  $\Delta y_i$ , то  $\hat{\beta}_j$  изменится на величину

$$\Delta \hat{\beta}_j = c_{ij} \Delta y_i.$$

Таким образом, в качестве показателя влиятельности в данном случае следует взять величину

$$c_{ij} = \frac{\Delta \hat{\beta}_i}{\Delta y_i}.$$

Вектор  $(c_{1j}, \dots, c_{nj})^T$ , отражающий влияние изменений в  $y$  на коэффициент  $\hat{\beta}_j$  называют  $DFBETAS_j$ .(???)

$$DFBETAS_j = ((X^T X)^{-1})_j X^T.$$

### Теорема о разбиении регрессоров

(Фриш-Ву-Лоуэлл) Теорема не имеет определенного названия, хотя и является очень важной в регрессионном анализе. Мы ввели название «Теорема о разбиении регрессоров» исключительно из соображений удобства.

Предположим, что мы разбили все регрессоры на 2 группы. Объединим регрессоры каждой группы в матрицы:  $X_1$  и  $X_2$ . Без потери общности, будем предполагать, что

$$X = [X_1, X_2].$$

Это блочное разбиение матрицы  $X$ .

Разобьем вектор коэффициентов МНК на части в соответствии с разбиением регрессоров:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}.$$

Нормальные уравнения  $X^T X \hat{\beta} = X^T y$  можно переписать в блочном виде

$$\begin{bmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{bmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} X_1^T y \\ X_2^T y \end{pmatrix},$$

Получим систему:

$$\begin{aligned} X_1^T X_1 \hat{\beta}_1 + X_1^T X_2 \hat{\beta}_2 &= X_1^T y, \\ X_2^T X_1 \hat{\beta}_1 + X_2^T X_2 \hat{\beta}_2 &= X_2^T y. \end{aligned}$$

Из первого уравнения

$$\hat{\beta}_1 = (X_1^T X_1)^{-1} X_1^T (y - X_2 \hat{\beta}_2).$$

Отсюда видно, что если мы вычислим  $\hat{\beta}_2$ , то довольно просто вычислить  $\hat{\beta}_1$  — это оценки МНК в регрессии  $y - X_2 \hat{\beta}_2$  по  $X_1$ . Этот результат легко понять: если  $\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$  дает минимум суммы квадратов остатков в полной регрессии, то  $\hat{\beta}_1$  должен

давать минимум при фиксированных остальных коэффициентах ( $\beta_1 = \hat{\beta}_1$ ).

Подставим выражение для  $\hat{\beta}_1$  во второе уравнение:

$$X_2^T X_1 (X_1^T X_1)^{-1} X_1^T (y - X_2 \hat{\beta}_2) + X_2^T X_2 \hat{\beta}_2 = X_2^T y.$$

Это уравнение можно преобразовать следующим образом:

$$\begin{aligned} X_2^T X_2 \hat{\beta}_2 - X_2^T X_1 (X_1^T X_1)^{-1} X_1^T X_2 \hat{\beta}_2 &= \\ &= X_2^T y - X_2^T X_1 (X_1^T X_1)^{-1} X_1^T y, \end{aligned}$$

$$X_2^T X_2 \hat{\beta}_2 - X_2^T P_1 X_2 \hat{\beta}_2 = X_2^T y - X_2^T P_1 y,$$

$$X_2^T M_1 X_2 \hat{\beta}_2 = X_2^T M_1 y,$$

где  $P_1$  и  $M_1$  — операторы проекции на пространства  $\mathcal{L}(X_1)$  и  $\mathcal{L}^\perp(X_1)$  соответственно:

$$P_1 = X_1 (X_1^T X_1)^{-1} X_1^T,$$

$$M_1 = I_n - P_1 = I_n - X_1 (X_1^T X_1)^{-1} X_1^T.$$

Найдем отсюда  $\hat{\beta}_2$ :

$$\hat{\beta}_2 = (X_2^T M_1 X_2)^{-1} X_2^T M_1 y.$$

**Теорема** (о разбиении регрессоров).

Если  $\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$  — оценки МНК в регрессии  $y$  по  $X = [X_1, X_2]$ , то  $\hat{\beta}_2$  — оценки МНК в регрессии  $M_1 y$  по  $M_1 X_2$ . Остатки в обеих регрессиях совпадают.

Доказательство:

Оценки МНК в регрессии  $M_1 y$  по  $M_1 X_2$  равны

$$\begin{aligned} \hat{\beta}_2 &= ((M_1 X_2)^T (M_1 X_2))^{-1} (M_1 X_2)^T (M_1 y) = \\ &= (X_2^T M_1^2 X_2)^{-1} X_2^T M_1^2 y = (X_2^T M_1 X_2)^{-1} X_2^T M_1 y, \end{aligned}$$

а это, как мы показали выше, и есть та часть вектора оценок в регрессии  $y$  по  $X$ , которая соответствует  $X_2$ .

Нам осталось доказать, что остатки в обеих регрессиях совпадут. Выше мы показали, что

$$\hat{\beta}_1 = (X_1^T X_1)^{-1} X_1^T (y - X_2 \hat{\beta}_2).$$

Поэтому остатки в регрессии  $y$  по  $X$  равны

$$\begin{aligned} e &= y - X_1 \hat{\beta}_1 - X_2 \hat{\beta}_2 = y - X_1 (X_1^T X_1)^{-1} X_1^T (y - X_2 \hat{\beta}_2) - X_2 \hat{\beta}_2 = \\ &= y - P_1 (y - X_2 \hat{\beta}_2) - X_2 \hat{\beta}_2 = M_1 y - M_1 X_2 \hat{\beta}_2. \quad \blacksquare \end{aligned}$$

Интерпретация:

$M_1 y$  — остатки из регрессии  $y$  по  $X_1$ .

$(M_1 X_2)_j$  — остатки из регрессии  $(X_2)_j$  по  $X_1$ , где  $(M_1 X_2)_j$  —  $j$ -й столбец матрицы  $M_1 X_2$ ,  $(X_2)_j$  —  $j$ -й столбец матрицы  $X_2$ .

Таким образом, мы строим регрессию  $y$  по  $X$ , но берем  $y$  и регрессоры предварительно «очищенные» от составляющей, лежащей в  $\mathcal{L}(X_1)$ . Другими словами, здесь мы применяем МНК к проекциям исходных данных на  $\mathcal{L}^\perp(X_1)$ .

Теорему особенно часто применяют к случаю, когда матрица  $X_1$  является вектором, состоящим из единиц:

$$X_1 = \mathbb{1}.$$

Матрица  $M_1$  в этом случае есть оператор центрирования. Для произвольного вектора  $z$  имеем

$$M_1 z = z - \mathbb{1}_n (\mathbb{1}_n^\top \mathbb{1}_n)^{-1} \mathbb{1}_n^\top z = z - \mathbb{1}_n \frac{1}{n} \mathbb{1}_n^\top z = z - \bar{z} \mathbb{1}_n = z^c.$$

Здесь  $\bar{z} = \frac{1}{n} \sum_i z_i = \frac{1}{n} \mathbb{1}_n^\top z$  — среднее значение переменной  $z$ ,  $z^c = z - \bar{z} \mathbb{1}_n$  — центрированный вектор  $z$ .

Таким образом,  $M_1 y$  — центрированная зависимая переменная,  $M_1 X_2$  — матрица составленная из центрированных регрессоров (всех регрессоров за исключением константы). Введем обозначения

$$y^c = M_1 y \quad \text{и} \quad X^c = M_1 X_2.$$

По теореме о разбиении регрессоров имеем

$$\hat{\beta}_2 = (X^{c\top} X^c)^{-1} X^{c\top} y^c,$$

Это оценки в регрессии  $y^c$  по  $X^c$ . По той же теореме остатки в центрированной регрессии совпадут с остатками в исходной регрессии. Вывод состоит в том, что центрирование переменных не влияет на результаты оценивания регрессии, если в регрессии содержится константа (понятно, что если применять центрирование, то константу приходится вычислять отдельно).

Применим данную теорему, чтобы получить формулы для расчета коэффициентов МНК в случае простой регрессии (два регрессора, один из которых — константа):

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i.$$

В прежних обозначениях

$$X_1 = \mathbb{1}_n, \quad X_2 = x.$$

Оценка  $\hat{\beta}_2$  находится как оценка в центрированной регрессии:

$$\hat{\beta}_2 = (\mathbf{x}^c \mathbf{x}^c)^{-1} \mathbf{x}^c \mathbf{y}^c = \frac{\mathbf{x}^c \mathbf{y}^c}{\mathbf{x}^c \mathbf{x}^c}.$$

Коэффициент  $\hat{\beta}_1$  можно найти через  $\hat{\beta}_2$ :

$$\hat{\beta}_1 = (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T (\mathbf{y} - \mathbf{X}_2 \hat{\beta}_2) = \frac{1}{n} \mathbf{1}_n^T (\mathbf{y} - \hat{\beta}_2 \mathbf{x}),$$

откуда

$$\hat{\beta}_1 = \bar{y} - \bar{x} \hat{\beta}_2.$$

Если  $\mathbf{x}$ ,  $\mathbf{y}$  — вектора, состоящие из независимых одинаково распределенных случайных величин, то несмещенная оценка их ковариации вычисляется по формуле:

$$\text{C}\hat{\text{ov}}(x, y) = \frac{1}{n-1} \mathbf{x}^c \mathbf{y}^c,$$

оценка дисперсий этих случайных величин — по формулам

$$\text{V}\hat{\text{ar}}(x) = \frac{1}{n-1} \mathbf{x}^c \mathbf{x}^c, \quad \text{V}\hat{\text{ar}}(y) = \frac{1}{n-1} \mathbf{y}^c \mathbf{y}^c,$$

а оценка коэффициента корреляции — по формуле

$$\hat{\rho}(x, y) = \text{C}\hat{\text{ov}}(x, y) / \sqrt{\text{V}\hat{\text{ar}}(x) \text{V}\hat{\text{ar}}(y)}.$$

Таким образом, в этом случае можно переписать формулу для коэффициента при  $\mathbf{x}$  следующим образом:

$$\hat{\beta}_2 = \frac{\mathbf{x}^c \mathbf{y}^c}{\mathbf{x}^c \mathbf{x}^c} = \frac{\text{C}\hat{\text{ov}}(x, y)}{\text{V}\hat{\text{ar}}(x)}$$

либо

$$\hat{\beta}_2 = \hat{\rho}(x, y) \cdot \sqrt{\frac{\text{V}\hat{\text{ar}}(y)}{\text{V}\hat{\text{ar}}(x)}}.$$

Центрируя переменные, мы «очищаем» их от среднего. Но эти же рассуждения применимы в случае, когда один из регрессоров является линейным трендом, то есть переменной, которая линейно меняется от номера наблюдения, например, переменной вида

$$\begin{pmatrix} 1 \\ 2 \\ 3 \\ \vdots \\ n \end{pmatrix}.$$

А именно, если в регрессии есть константа и тренд, то оценки МНК должны совпасть с оценками, полученными из регрессии, в которой все переменные «очищены» от среднего и

тренда. Для этого частного случая и была впервые Р. Фришем сформулирована данная теорема.

### Вычисление оценок МНК и других регрессионных величин

Непосредственный метод

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

— не самый быстрый и простой. Требуется обращать матрицу, либо решать систему уравнений

$$(X^T X)\hat{\beta} = X^T y.$$

Метод триангуляризации:

$$X^T X = T^T T,$$

где  $T$  — верхняя (или же нижняя) треугольная матрица. Такое представление называют разложением Холецкого или триангуляризацией. Оно существует для любой симметричной положительно полуопределенной матрицы, т.е. и для  $X^T X$ . Не сложно построить рекуррентный алгоритм для расчета  $T$  по данной матрице  $X^T X$ . Мы не будем давать здесь формальное описание алгоритма, приведем лишь конкретный пример, из которого понятно, как алгоритм работает.

Пример.

$$X = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 5 & 3 \\ 1 & 0 & 4 \\ 1 & 5 & 6 \end{bmatrix}, \quad X^T X = \begin{bmatrix} 4 & 10 & 14 \\ 10 & 50 & 45 \\ 14 & 45 & 62 \end{bmatrix}.$$

$$T = \begin{bmatrix} t_{11} & t_{12} & t_{13} \\ 0 & t_{22} & t_{23} \\ 0 & 0 & t_{33} \end{bmatrix},$$

$$T^T T = \begin{bmatrix} t_{11}^2 & t_{11}t_{12} & t_{11}t_{13} \\ t_{11}t_{12} & t_{12}^2 + t_{22}^2 & t_{12}t_{13} + t_{22}t_{23} \\ t_{11}t_{13} & t_{12}t_{13} + t_{22}t_{23} & t_{13}^2 + t_{23}^2 + t_{33}^2 \end{bmatrix},$$

$$t_{11} = \sqrt{4} = 2,$$

$$t_{12} = 10/t_{11} = 5, \quad t_{22} = \sqrt{50 - t_{12}^2} = 5,$$

$$t_{13} = 14/t_{11} = 7, \quad t_{23} = (45 - t_{12}t_{13})/t_{22} = 10/5 = 2,$$

$$t_{33} = \sqrt{62 - t_{13}^2 - t_{23}^2} = \sqrt{62 - 49 - 4} = 3.$$

$$T = \begin{bmatrix} 2 & 5 & 7 \\ 0 & 5 & 2 \\ 0 & 0 & 3 \end{bmatrix}.$$

(Диагональные элементы матрицы  $T$  для определенности везде выбираем положительными.)

Уравнения для нахождения  $\hat{\beta}$  примут вид

$$T^T T \hat{\beta} = X^T y.$$

Эту систему уравнений можно решать в два этапа. Сначала находим вектор  $t$  из системы уравнений

$$T^T t = X^T y,$$

где  $t = (T^T)^{-1} X^T y$ . Потом находим  $\hat{\beta}$  из системы уравнений

$$T \hat{\beta} = t.$$

Каждый раз решаем систему уравнений с треугольной матрицей, а это довольно просто.

#### Пример.

Пусть матрица  $X$  та же. Возьмем

$$y = \begin{pmatrix} 1 \\ 4 \\ -2 \\ 1 \end{pmatrix}.$$

Тогда

$$X^T y = \begin{pmatrix} 4 \\ 25 \\ 11 \end{pmatrix}.$$

Найдем  $t$  из системы  $T^T t = X^T y$ :

$$\begin{aligned} 2t_1 &= 4 \\ 5t_1 + 5t_2 &= 25 \\ 7t_1 + 2t_2 + 3t_3 &= 11. \end{aligned}$$

Откуда

$$\begin{aligned} t_1 &= 4/2 = 2, \\ t_2 &= (25 - 5 \cdot 2)/5 = 3, \\ t_3 &= (11 - 7 \cdot 2 - 2 \cdot 3)/3 = -3. \end{aligned}$$

Теперь найдем  $\hat{\beta}$  из системы  $T \hat{\beta} = t$ :

$$\begin{aligned} 2\hat{\beta}_1 + 5\hat{\beta}_2 + 7\hat{\beta}_3 &= 2 \\ 5\hat{\beta}_2 + 2\hat{\beta}_3 &= 3 \\ 3\hat{\beta}_3 &= -3. \end{aligned}$$

Откуда

$$\begin{aligned} \hat{\beta}_3 &= -1, \\ \hat{\beta}_2 &= (3 - 2 \cdot (-1))/5 = 1, \\ \hat{\beta}_1 &= (2 - 5 \cdot 1 - 7 \cdot (-1))/2 = 2. \end{aligned}$$

Обозначим  $\check{X} = X T^{-1}$ .<sup>1</sup> Поскольку  $\check{X}$  получается из  $X$  линейным преобразованием, то подпространства, натянутые на столбцы этих матриц совпадают:

$$\mathcal{L}(\check{X}) = \mathcal{L}(X).$$

Кроме того, матрица  $\check{X}$  имеет ортонормированные столбцы:

$$\check{X}^T \check{X} = (T^T)^{-1} X^T X T^{-1} = (T^T)^{-1} T^T T T^{-1} = I_m.$$

Таким образом, столбцы матрицы  $\check{X}$  составляют ортонормированный базис подпространства  $\mathcal{L}(X)$ .

Кроме того, отметим, что

$$t = (T^T)^{-1} X^T y = \check{X}^T y.$$

Практически все алгоритмы нахождения оценок МНК построены на этом принципе.

Есть простой способ получения ортонормированного базиса подпространства, натянутого на некоторую систему векторов — [ортогонализация Грама-Шмидта](#). Этот алгоритм построен так, что одновременно мы получаем коэффициенты преобразования в виде треугольной матрицы, а это нам и требуется в преобразовании Холецкого.<sup>2</sup> (Матрица ортонормирующего преобразования будет верхней треугольной.)

Алгоритм ортогонализации Грама-Шмидта состоит из двух простых операций. Первая операция — нормирование. Под нормированным вектором  $\check{x}$  мы понимаем здесь вектор  $\check{x}$  единичной длины ( $\|\check{x}\| = 1$ ), который получается из  $x$  умножением на константу:

$$\check{x} = \frac{x}{\|x\|}.$$

Вторая операция — получение из вектора  $z$  вектора  $\tilde{z}$ , который был бы ортогонален другому вектору —  $x$  (ортогонализация). Нам требуется найти число  $\alpha$ , такое чтобы для

<sup>1</sup> Такое представление матрицы принято называть QR-разложением. В данном случае матрица  $\check{X}$  играет роль матрицы  $Q$ , которая должна быть ортогональной, а  $T^{-1}$  играет роль матрицы  $R$ , которая должна быть верхней треугольной.

<sup>2</sup> Метод ортогонализации Грама-Шмидта менее подвержен ошибкам округления, чем метод, основанный непосредственно на разложении Холецкого матрицы  $X^T X$  (пример этого алгоритма был приведен выше).

вектора  $\tilde{z} = z - \alpha x$  выполнялось  $\tilde{z}^\top x = 0$ . Из уравнения  $(y - \alpha x)^\top x = 0$  найдем  $\alpha = z^\top x / x^\top x$ , откуда

$$\tilde{z} = z - \frac{z^\top x}{x^\top x} x.$$

Пусть требуется ортонормировать набор регрессоров  $(x_1, \dots, x_n)$  — получить из него ортонормированный набор  $(\check{x}_1, \dots, \check{x}_n)$ .

Первый вектор получить несложно — достаточно нормировать  $x_1$ :

$$\check{x}_1 = \frac{x_1}{\|x_1\|}.$$

Рассмотрим шаг рекуррентного алгоритма ортогонализации. Пусть мы получили первые  $k-1$  векторов:  $(\check{x}_1, \dots, \check{x}_{k-1})$ . Найдем на основе  $x_k$  ортогональный им вектор  $\tilde{x}_k$  из  $\mathfrak{L}(X)$ . Его можно получить следующим образом. Обозначим

$$t_{jk} = \check{x}_j^\top x_k \quad (j = 1, \dots, k-1).$$

Тогда

$$\tilde{x}_k = x_k - \sum_{j=1}^{k-1} t_{jk} \check{x}_j.$$

Этот вектор, как несложно проверить, действительно ортогонален полученным ранее:

$$\check{x}_s^\top \tilde{x}_k = \check{x}_s^\top x_k - \sum_{j=1}^{k-1} t_{jk} \check{x}_s^\top \check{x}_j = t_{sk} - t_{sk} = 0.$$

Далее вектор  $\tilde{x}_k$  следует нормировать:

$$\check{x}_k = \frac{\tilde{x}_k}{\|\tilde{x}_k\|}.$$

Можно обозначить  $t_{kk} = \|\tilde{x}_k\|$ , при этом  $\check{x}_k = t_{kk}^{-1} \tilde{x}_k$ , и можно записать следующую формулу:

$$x_k = \sum_{j=1}^k t_{jk} \check{x}_j \quad (k = 2, \dots, m).$$

Чтобы эта формула была верна при  $k = 1$ , обозначим  $t_{11} = \|x_1\|$ .

Таким образом, мы получили требуемое:

$$X = T \check{X},$$

где  $T$  — верхняя треугольная матрица, составленная из вычисленных в результате алгоритма чисел  $t_{jk}$ :

$$T = \begin{bmatrix} t_{11} & \cdots & t_{1m} \\ \vdots & \ddots & \vdots \\ \textcircled{0} & & t_{mm} \end{bmatrix}.$$

Как мы отмечали выше,  $t = \check{X}^T y$ , поэтому  
 $t_j = \check{x}_j^T y$ .

Имея вектор  $t$  и матрицу  $T$ , легко получить оценки  $\hat{\beta}$ :  
 $T \hat{\beta} = t$ .

Остатки можно вычислить по формуле

$$e = y - \check{X}t = y - \sum_{j=1}^m t_j \check{x}_j,$$

т.е. как дополнительные «полшага» того же алгоритма ортогонализации.<sup>3</sup>

Пример.

$$\begin{aligned} X &= \begin{bmatrix} 1 & 0 & 1 \\ 1 & 5 & 3 \\ 1 & 0 & 4 \\ 1 & 5 & 6 \end{bmatrix}, & y &= \begin{pmatrix} 3 \\ 18 \\ 2 \\ 13 \end{pmatrix}, \\ t_{11} &= \|x_1\| = \sqrt{4} = 2, & \check{x}_1 &= \begin{pmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{pmatrix}, \\ t_{12} &= 1/2 \cdot 0 + 1/2 \cdot 5 + 1/2 \cdot 0 + 1/2 \cdot 5 = 5. \\ \tilde{x}_2 &= x_2 - t_{21} \check{x}_1 = \begin{pmatrix} -5/2 \\ 5/2 \\ -5/2 \\ 5/2 \end{pmatrix}, \\ t_{22} &= \|\tilde{x}_2\| = \sqrt{\frac{25}{4} \cdot 4} = 5, & \check{x}_2 &= \begin{pmatrix} -1/2 \\ 1/2 \\ -1/2 \\ 1/2 \end{pmatrix}, \\ t_{13} &= 1/2 \cdot 1 + 1/2 \cdot 3 + 1/2 \cdot 4 + 1/2 \cdot 6 = 7, \\ t_{23} &= -1/2 \cdot 1 + 1/2 \cdot 3 - 1/2 \cdot 4 + 1/2 \cdot 6 = 2, \\ \tilde{x}_3 &= x_3 - t_{31} \check{x}_1 - t_{32} \check{x}_2 = \begin{pmatrix} -3/2 \\ -3/2 \\ 3/2 \\ 3/2 \end{pmatrix}, \end{aligned}$$

<sup>3</sup> Можно на основе остатков вычислить  $t$  как  $t = \check{X}^T e$ . Эта формула менее подвержена ошибкам округления.

$$t_{33} = \|\tilde{x}_3\| = \sqrt{\frac{9}{4} \cdot 4} = 3, \quad \check{x}_3 = \begin{pmatrix} -1/2 \\ -1/2 \\ 1/2 \\ 1/2 \end{pmatrix},$$

$$T = \begin{bmatrix} 2 & 5 & 7 \\ 0 & 5 & 2 \\ 0 & 0 & 3 \end{bmatrix},$$

$$t_1 = \check{x}_1^T y = 1/2 \cdot 3 + 1/2 \cdot 18 + 1/2 \cdot 2 + 1/2 \cdot 13 = 18,$$

$$t_2 = \check{x}_2^T y = -1/2 \cdot 3 + 1/2 \cdot 18 - 1/2 \cdot 2 + 1/2 \cdot 13 = 13,$$

$$t_3 = \check{x}_3^T y = -1/2 \cdot 3 + 1/2 \cdot 18 - 1/2 \cdot 2 + 1/2 \cdot 13 = -3,$$

$$e = y - t_1 \check{x}_1 - t_2 \check{x}_2 - t_3 \check{x}_3.$$

$$e_1 = 3 - 1/2 \cdot 18 - (-1/2) \cdot 13 - (-1/2) \cdot (-3) = -1.$$

$$e_2 = 18 - 1/2 \cdot 18 - 1/2 \cdot 13 - (-1/2) \cdot (-3) = 1.$$

$$e_3 = 2 - 1/2 \cdot 18 - (-1/2) \cdot 13 - 1/2 \cdot (-3) = 1.$$

$$e_4 = 13 - 1/2 \cdot 18 - 1/2 \cdot 13 - 1/2 \cdot (-3) = -1.$$

$$\text{Решая уравнения } T \hat{\beta} = t, \text{ находим } \hat{\beta} = \begin{pmatrix} 5 \\ 3 \\ -1 \end{pmatrix}.$$

### Ортогональная матрица регрессоров

Пусть матрица  $X$  такова, что регрессоры попарно ортогональны, т.е. для любой пары регрессоров  $x_j, x_s$  ( $j \neq s$ ) выполняется  $x_j^T x_s = 0$ . При этом матрица  $X^T X$  будет диагональной:

$$X^T X = \text{diag}(x_1^T x_1, \dots, x_m^T x_m).$$

Тогда

$$(X^T X)^{-1} = \text{diag}\left(\frac{1}{x_1^T x_1}, \dots, \frac{1}{x_m^T x_m}\right),$$

$$\hat{\beta} = \text{diag}\left(\frac{1}{x_1^T x_1}, \dots, \frac{1}{x_m^T x_m}\right) X^T y,$$

т.е.

$$\hat{\beta}_j = \frac{x_j^T y}{x_j^T x_j}.$$

Это соотношение показывает, что оценка коэффициента при любом регрессоре не зависит от величины любого другого регрессора. Кроме того, если часть регрессоров отбросить, то оценки коэффициентов при оставшихся регрессорах не изменятся, т.е. если  $X = [X_1, X_2]$ , и

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = (X^T X)^{-1} \begin{pmatrix} X_1^T y \\ X_2^T y \end{pmatrix},$$

то

$$\hat{\beta}_1 = (X_1^T X_1)^{-1} X_1^T y \quad (\text{и} \quad \hat{\beta}_2 = (X_2^T X_2)^{-1} X_2^T y).$$

В случае ортогональных регрессоров можно считать, что в представлении

$$y = \hat{\beta}_1 x_1 + \dots + \hat{\beta}_m x_m + e$$

слагаемое  $\hat{\beta}_j x_j$  — это та составляющая вектора  $y$ , которая «объяснена»  $j$ -м регрессором. Считая так, мы не столкнемся с затруднениями, поскольку все слагаемые здесь ортогональны между собой:

$$x_j^T x_s = 0 \quad \text{при} \quad j \neq s$$

согласно условиям, а

$$x_j^T e = 0$$

— нормальные уравнения, выполняющиеся для оценок МНК.

Поэтому в произведении  $y^T y$  все перекрестные члены должны «занулиться»:

$$y^T y = \hat{\beta}_1^2 x_1^T x_1 + \dots + \hat{\beta}_m^2 x_m^T x_m + e^T e.$$

Здесь

$y^T y$  — общая сумма квадратов ( $TSS$ ),

$\hat{\beta}_j^2 x_j^T x_j$  — то, что «объяснено» в общей сумме квадратов  $j$ -м регрессором ( $ESS_j$ ),

$e^T e$  — остаточная («необъясненная») сумма квадратов ( $RSS$ ).

Таким образом, получим следующее разложение суммы квадратов:

$$TSS = \sum_j ESS_j + RSS.$$

Можно рассмотреть более общий случай. Пусть  $X$  можно разбить на  $k$  попарно ортогональных подматриц:

$$X = [X_1, \dots, X_k],$$

и

$$X_s^T X_t = 0 \quad (s \neq t).$$

Тогда матрица  $X^T X$  будет блочно-диагональной:

$$\mathbf{X}^\top \mathbf{X} = \begin{bmatrix} \mathbf{X}_1^\top \mathbf{X}_1 & \cdots & \mathbb{O} \\ \mathbb{O} & \cdots & \mathbf{X}_k^\top \mathbf{X}_k \end{bmatrix}.$$

Отсюда

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \begin{bmatrix} (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} & \cdots & \mathbb{O} \\ \mathbb{O} & \cdots & (\mathbf{X}_k^\top \mathbf{X}_k)^{-1} \end{bmatrix},$$

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} & \cdots & \mathbb{O} \\ \mathbb{O} & \cdots & (\mathbf{X}_k^\top \mathbf{X}_k)^{-1} \end{bmatrix} \begin{pmatrix} \mathbf{X}_1^\top \mathbf{y} \\ \vdots \\ \mathbf{X}_k^\top \mathbf{y} \end{pmatrix} = \begin{pmatrix} (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top \mathbf{y} \\ \vdots \\ (\mathbf{X}_k^\top \mathbf{X}_k)^{-1} \mathbf{X}_k^\top \mathbf{y} \end{pmatrix}.$$

Таким образом, коэффициенты МНК можно рассчитывать отдельно для каждой из подматриц:

$$\hat{\boldsymbol{\beta}}_s = (\mathbf{X}_s^\top \mathbf{X}_s)^{-1} \mathbf{X}_s^\top \mathbf{y}.$$

На эти оценки никак не влияет то, каковы другие подматрицы  $\mathbf{X}_t$  ( $t \neq s$ ) и присутствуют они в регрессии или нет.

Поскольку

$$\mathbf{y} = \hat{\boldsymbol{\beta}}_1 \mathbf{X}_1 + \dots + \hat{\boldsymbol{\beta}}_k \mathbf{X}_k + \mathbf{e},$$

то

$$\mathbf{y}^\top \mathbf{y} = \hat{\boldsymbol{\beta}}_1^\top \mathbf{X}_1^\top \mathbf{X}_1 \hat{\boldsymbol{\beta}}_1 + \dots + \hat{\boldsymbol{\beta}}_k^\top \mathbf{X}_k^\top \mathbf{X}_k \hat{\boldsymbol{\beta}}_k + \mathbf{e}^\top \mathbf{e}.$$

Здесь  $\hat{\boldsymbol{\beta}}_s^\top \mathbf{X}_s^\top \mathbf{X}_s \hat{\boldsymbol{\beta}}_s$  — сумма квадратов, объясненная  $s$ -й группой регрессоров.

### Линейные преобразования переменных регрессии

Пусть переменные регрессии были преобразованы следующим образом:

$$\tilde{\mathbf{y}} = \alpha \mathbf{y} + \mathbf{X} \mathbf{a}, \quad \tilde{\mathbf{X}} = \mathbf{X} \mathbf{A},$$

где  $\alpha$  — скаляр,  $\mathbf{a}$  — вектор,  $\mathbf{A}$  — квадратная невырожденная матрица.

Оценки МНК в «преобразованной» регрессии равны

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &= (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \tilde{\mathbf{y}} = (\mathbf{A}^\top \mathbf{X}^\top \mathbf{X} \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{X}^\top (\alpha \mathbf{y} + \mathbf{X} \mathbf{a}) = \\ &= (\mathbf{A}^\top)^{-1} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\alpha \mathbf{y} + \mathbf{X} \mathbf{a}) = \\ &= (\mathbf{A}^\top)^{-1} (\alpha (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} \mathbf{a}), \end{aligned}$$

т.е.

$$\tilde{\boldsymbol{\beta}} = (\mathbf{A}^\top)^{-1} (\alpha \hat{\boldsymbol{\beta}} + \mathbf{a}).$$

Таким образом, линейные преобразования переменных приводят к линейным преобразованиям коэффициентов.

Можно восстановить коэффициенты МНК исходной регрессии:

$$\hat{\beta} = \frac{1}{\alpha}(A^T \hat{\beta} - a).$$

Матрица проекции на  $\mathcal{L}^\perp(\tilde{X})$  совпадает с матрицей проекции на  $\mathcal{L}^\perp(X)$ :

$$\begin{aligned}\tilde{M} &= I - \tilde{X}(\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T = I - X A (A^T X^T X A)^{-1} A^T X^T \\ &= I - X (X^T X)^{-1} X^T = M.\end{aligned}$$

Из этого следует, что остатки в «преобразованной» регрессии равны

$$\tilde{e} = \tilde{M} \tilde{y} = M \tilde{y} = \alpha M y + M X a = \alpha M y = \alpha e.$$

На остатки, таким образом, влияет только  $\alpha$  (множитель при  $y$ ), и если  $\alpha = 1$ , то остатки совпадут с остатками в исходной регрессии.

## Линейная регрессия как вероятностная модель. Свойства оценок

Процесс, который порождает данные:

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

В этой модели  $\epsilon$  — ненаблюдаемые случайные величины, называемые ошибками или возмущениями,  $y$  и  $X$  известны, а  $\beta$  — вектор неизвестных параметров. Предполагается, что был некоторый вектор истинных параметров  $\hat{\beta}$ , т.е. данные были порождены указанным процессом при  $\beta = \hat{\beta}$ , и требуется каким-то образом оценить этот вектор на основании имеющихся данных ( $y$  и  $X$ ).

При этом делаются какие-то предположения о том, какое распределение имеют ошибки  $\epsilon$  и матрица регрессоров  $X$ . Имеются две разновидности вероятностных моделей регрессии:

- 1) регрессоры — детерминированные величины (не случайные);
- 2) регрессоры — случайные величины (по крайней мере часть из них).

В дальнейшем мы будем по умолчанию предполагать, что имеем дело со случайными регрессорами и специально оговаривать, когда будем предполагать детерминированность регрессоров.

Об ошибках предполагают практически всегда, что они имеют нулевое мат. ожидание:

$$E(\boldsymbol{\varepsilon}) = \mathbf{0}.$$

(Здесь и в дальнейшем мат. ожидания берутся в предположении  $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ .)

Кроме того, часто предполагают, что  $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \sigma^2 \mathbf{I}_n$ . Поскольку  $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ , то  $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T)$  есть ковариационная матрица ошибок, т.е.

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \text{Var}(\boldsymbol{\varepsilon}).$$

Предположение  $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$  означает, во-первых, что ошибки разных наблюдений некоррелированы между собой:

$$\text{Cov}(\varepsilon_{i_1}, \varepsilon_{i_2}) = 0 \quad (i_1 \neq i_2)$$

При выполнении этого условия говорят, что отсутствует автокорреляция ошибок.

Во-вторых,  $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$  означает, что дисперсии ошибок для всех наблюдений одинаковы:

$$\text{Var}(\varepsilon_i) = \sigma^2 \quad \forall i.$$

Если это условие выполнено, то говорят, что ошибка гомоскедастична, а если не выполнено, то говорят, что ошибка гетероскедастична.

Как правило, при оценивании регрессии истинная величина дисперсии ошибки  $\sigma^2$ , (которую, по аналогии с истинной величиной коэффициентов обозначим  $\hat{\sigma}^2$ ) неизвестна, то есть  $\sigma^2$  является неизвестным параметром, который требуется оценить наряду с коэффициентами  $\boldsymbol{\beta}$ .

Процедура оценивания, которую мы используем, даст вектор оценок  $\hat{\boldsymbol{\beta}}$ . Процедуру (способ) оценивания, как правило, можно представить как функцию  $\hat{\boldsymbol{\beta}}(y, X)$ , которая показывает, как по имеющимся данным получить оценки. Такую функцию называют оценивателем. (Англ. *estimator* — оцениватель в отличие от *estimate* — оценка. К сожалению в русскоязычной литературе вместо оценивателя говорят об оценках. Мы в дальнейшем тоже будем пользоваться термином «оценки», если это не приводит к путанице.) Заметим,

что оценщик должен быть определен для матриц с различным количеством столбцов, что соответствует разному количеству имеющихся наблюдений  $n$ .

Оценки наименьших квадратов, как мы видели в предыдущей главе, задаются формулой

$$\hat{\beta}_{\text{МНК}}(y, X) = (X^T X)^{-1} X^T y.$$

Есть и другие способы оценивания. Но в дальнейшем мы сосредоточим внимание на МНК, как самом удобном способе оценивания, обладающим, к тому же рядом «хороших» свойств.

Заметим, что вектор оценок  $\hat{\beta}$  получают на основе случайных данных, поэтому это случайная величина.

Чего мы хотим от оценок? Должна быть некоторая целевая функция.

Обычно рассматривают функцию потерь  $L(\beta, \hat{\beta})$  — какие потери мы несем, если получили оценку  $\hat{\beta}$  при истинном значении параметров  $\beta$ . Хотелось бы найти такой оценщик  $\hat{\beta}(\cdot, \cdot)$ , чтобы он минимизировал ожидаемые потери, т.е.

$$E(L(\hat{\beta}(y, X), \hat{\beta})) \rightarrow \min_{\hat{\beta}(\cdot, \cdot)}.$$

Заметьте, что это оптимизация по функции, так как оценщик — функция данных.

Самой распространенной и простой является квадратичная функция потерь:

$$L(\beta, \hat{\beta}) = (\beta - \hat{\beta})^T Q (\beta - \hat{\beta}),$$

где  $Q$  — некоторая симметричная положительно определенная матрица. Это просто обобщение евклидова расстояния между  $\beta$  и  $\hat{\beta}$ .

(Разложение в ряд Тейлора + симметричность)

Ожидаемые потери при квадратичной функции потерь называют среднеквадратической ошибкой (или средним квадратом ошибки; здесь слово «средний» используется в смысле «ожидаемый»):

$$\text{MSE } Q(\hat{\beta}) = E((\hat{\beta} - \beta)^T Q (\hat{\beta} - \beta)).$$

Ошибку оценивания можно представить как сумму двух компонент:

$$\hat{\beta} - \beta = [\hat{\beta} - E(\hat{\beta})] + [E(\hat{\beta}) - \beta].$$

Здесь первое слагаемое имеет нулевое мат. ожидание:

$$E(\hat{\beta} - E(\hat{\beta})) = 0.$$

Второе слагаемое называют **смещением** или систематической ошибкой:

$$\text{Bias}(\hat{\beta}) = E(\hat{\beta}) - \beta.$$

Поскольку смещение — детерминированная величина, то

$$\begin{aligned} E[(\hat{\beta} - E(\hat{\beta}))^T Q (E(\hat{\beta}) - \beta)] &= E[\hat{\beta} - E(\hat{\beta})]^T Q (E(\hat{\beta}) - \beta) \\ &= 0^T Q (E(\hat{\beta}) - \beta) = 0. \end{aligned}$$

Отсюда следует, что среднеквадратическую ошибку тоже можно представить как сумму двух компонент:

$$\begin{aligned} \text{MSE } Q(\hat{\beta}) &= \\ &= E((\hat{\beta} - E(\hat{\beta})) + [E(\hat{\beta}) - \beta])^T Q ((\hat{\beta} - E(\hat{\beta})) + [E(\hat{\beta}) - \beta]) = \\ &= E((\hat{\beta} - E(\hat{\beta}))^T Q (\hat{\beta} - E(\hat{\beta}))) + (E(\hat{\beta}) - \beta)^T Q (E(\hat{\beta}) - \beta). \end{aligned}$$

Удобно переписать среднеквадратическую ошибку в следующем виде:

$$\begin{aligned} \text{MSE } Q(\hat{\beta}) &= \\ &= E(\text{tr}[(\hat{\beta} - E(\hat{\beta}))^T Q (\hat{\beta} - E(\hat{\beta}))]) + \text{Bias}(\hat{\beta})^T Q \text{Bias}(\hat{\beta}) = \\ &= \text{tr}[E((\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))^T) Q] + \text{Bias}(\hat{\beta})^T Q \text{Bias}(\hat{\beta}) = \\ &= \text{tr}[\text{Var}(\hat{\beta}) Q] + \text{Bias}(\hat{\beta})^T Q \text{Bias}(\hat{\beta}). \end{aligned}$$

При выводе этой формулы мы использовали стандартный прием, основанный на двух простых фактах:

- (•) След скалярной величины есть сама эта скалярная величина.
- (•) След произведения матриц не меняется при перестановке матриц:

$$\text{tr}(AB) = \text{tr}(BA).$$

Первое из слагаемых данного разложения соответствует ковариационной матрице оценок ( $\text{Var}(\hat{\beta})$ ) и служит мерой «разброса» оценок. Второе из слагаемых соответствует смещению оценок.

К сожалению, невозможно напрямую минимизировать **MSE** (или любые другие ожидаемые потери). Эта величина зависит от неизвестных характеристик распределения данных, в том числе от неизвестных истинных параметров  $\beta$ . Эта проблема может быть решена в рамках **байесовского подхода** к оцениванию. Согласно этому подходу, параметры  $\beta$  являются случайной величиной, о распределении которой нам может быть что-то известно (другими словами, у нас имеется так называемая **априорная информация**). Тогда можно рассчи-

татъ среднеквадратическую ошибку условную по той информации, которая у нас имеется. Хотя этот подход идейно привлекателен, но он более сложен, чем традиционная статистическая теория, и мы не будем им заниматься. Достаточно сказать, что в рамках байесовского подхода при определенных предположениях оценки МНК являются оптимальными с точки зрения квадратичной функции потерь.

В традиционной статистике используют не столь утилитарный подход. Обычно просто требуют, чтобы оценки обладали некоторыми «хорошими» свойствами.

### Свойства оценок МНК в конечных выборках

Одним из «хороших» свойств оценок является несмещенность. Согласно определению, оценка является несмещенной, когда

$$\text{Bias}(\hat{\beta}) = 0,$$

т.е.

$$E(\hat{\beta}) = \beta.$$

#### Теорема.

Если  $E(\epsilon|X) = 0$ , то оценки МНК являются несмещенными.

#### Замечание.

Если матрица  $X$  детерминированная, то

$$E(\epsilon|X) = 0 \Leftrightarrow E(\epsilon) = 0,$$

поэтому в этом случае в условиях теоремы достаточно было бы потребовать, чтобы безусловное мат. ожидание ошибок было нулевым:

$$E(\epsilon) = 0.$$

Предположение  $E(\epsilon|X) = 0$ , можно условно назвать «ортogonalность между регрессорами и ошибками». Это ортogonalность в довольно сильном смысле:

$E(\epsilon|X) = 0$  выполнено тогда и только тогда, когда выполнено  $E(F(X)\epsilon) = 0$  для любой матричной функции  $F(X) : \mathbb{R}^{m \times n} \mapsto \mathbb{R}^n$  (подходящей размерности) ????

Т.е. данное условие эквивалентно тому, что ошибки в вероятностном смысле ортогональны любой функции регрессоров.

Условие  $E(\epsilon|X) = 0$  можно разложить на две части:

(1)  $E(\epsilon) = 0$  (мат. ожидание ошибки равно нулю);

(2)  $E(\epsilon|X) = E(\epsilon)$  (условное по  $X$  мат. ожидание ошибки равно безусловному).

Второе свойство есть нечто среднее между независимостью и некоррелированностью (Из независимости  $\epsilon$  и  $X$  следует свойство (2), а из свойства (2) следует некоррелированность  $\epsilon$  и  $X$ ). Притом это асимметричная «независимость» — мат. ожидание  $\epsilon$  не зависит от  $X$ , но не наоборот.

Доказательство:

Покажем прежде всего, что оценки МНК  $\hat{\beta}$  являются несмещенными условно по  $X$ .

$$\begin{aligned} E(\hat{\beta}|X) &= E((X^T X)^{-1} X^T y|X) = (X^T X)^{-1} X^T E(y|X) = \\ &= (X^T X)^{-1} X^T E(X\hat{\beta} + \epsilon|X) = \hat{\beta} + (X^T X)^{-1} X^T E(\epsilon|X) = \hat{\beta}. \end{aligned}$$

По правилу полного мат. ожидания  $E(\hat{\beta}) = E(E(\hat{\beta}|X))$ , поэтому

$$E(\hat{\beta}) = E(\hat{\beta}) = \hat{\beta}. \quad \blacksquare$$

Заметим, что несмещенная оценка не всегда самая хорошая с точки зрения ожидаемых потерь. Например, если использовать квадратичную функцию потерь, то важна также дисперсия. В статистике не столь редки случаи, когда несмещенная оценка имеет слишком большую дисперсию, и поэтому использовать ее нежелательно.

Если потребовать выполнение достаточно сильных условий, в том числе рассматривать только оценки из определенного класса, то оценитель, минимизирующий  $MSE Q$ , не будет зависеть от неизвестных параметров. Оказывается, что оценитель наименьших квадратов является оптимальным в классе линейных несмещенных оценок (в англоязычной литературе для обозначения этого свойства используют аббревиатуру BLUE — best linear unbiased estimator). Поскольку рассматриваются несмещенные оценки, то слагаемое в разложении  $MSE Q$ , относящееся к смещению, равно нулю и остается

только слагаемое, относящееся к дисперсии. То, что оценки МНК имеют наименьшую дисперсию (такие оценки называют эффективными) в классе линейных несмещенных оценок, эквивалентно тому, что они имеют минимальную величину  $MSE Q$  среди линейных несмещенных оценок.

**Теорема** (Гаусса-Маркова).

Пусть

(а) матрица регрессоров  $X$  детерминирована,

(б)  $E(\epsilon) = 0$ ,

(в)  $E(\epsilon\epsilon^T) = \sigma^2 I$ .

Тогда оценки МНК  $\hat{\beta} = (X^T X)^{-1} X^T y$  имеют наименьшую дисперсию в классе линейных по  $y$  несмещенных оценок.

Замечание.

Определенные трудности возникают с тем, в каком смысле можно понимать дисперсию векторной случайной величины. Ведь аналогом дисперсии в данном случае является ковариационная матрица  $\text{Var}(\hat{\beta})$ .

Пусть  $A$  и  $B$  — две положительно полуопределенные матрицы. Какая из них больше? Возможное определение состоит в следующем:  $A > B$ , если существует положительно определенная матрица  $\Psi$ , такая что  $A = B + \Psi$ ;  $A \geq B$ , если существует положительно полуопределенная матрица  $\Psi$ , такая что  $A = B + \Psi$ . Конечно, этот способ далеко не всегда позволяет сравнить две произвольные положительно полуопределенные матрицы. Но в данной теореме требуемая матрица  $\Psi$  существует, и поэтому сравнение возможно.

Другой вариант заключается в том, чтобы рассматривать не дисперсию вектора, а дисперсию линейной комбинации элементов этого вектора. При этом получаем следующее определение:

Дисперсия векторной случайной величины  $a$  больше, чем дисперсия векторной случайной величины  $b$  тогда и только тогда, когда  $\text{Var}(\gamma^T a) > \text{Var}(\gamma^T b)$  при любом ненулевом векторе  $\gamma$ .

Можно также следовать приведенной выше логике, и исходить из среднеквадратической ошибки. Мы можем рас-

смотреть обобщение дисперсии, которое уже является скалярной величиной:

$$\text{Var}_Q(\hat{\beta}) = E((\hat{\beta} - E(\hat{\beta}))^T Q (\hat{\beta} - E(\hat{\beta}))),$$

где  $Q$  — некоторая симметричная положительно определенная матрица. Как мы видели выше,

$$\text{Var}_Q(\hat{\beta}) = \text{tr}[\text{Var}(\hat{\beta}) Q].$$

(Мы использовали здесь, что для несмещенной оценки  $\text{MSE}_Q(\hat{\beta}) = \text{Var}_Q(\hat{\beta})$ )

Таким образом, мы можем доказывать одно из трех соотношений:

(1)  $\text{Var}(\hat{\beta}) = \text{Var}(\hat{\beta}_{\text{МНК}}) + \Psi$ , где  $\Psi$  — положительно полуопределенная матрица;

(2)  $\text{Var}(\gamma^T \hat{\beta}) \geq \text{Var}(\gamma^T \hat{\beta}_{\text{МНК}})$ ;

либо

(3)  $\text{Var}_Q(\hat{\beta}) \geq \text{Var}_Q(\hat{\beta}_{\text{МНК}})$ ,

где  $\hat{\beta}$  — произвольная линейная несмещенная оценка коэффициентов регрессии,  $\hat{\beta}_{\text{МНК}}$  — оценка, полученная методом наименьших квадратов.

Оказывается, все эти три интерпретации теоремы Гаусса-Маркова эквивалентны.

*Доказательство:*

Линейный по  $y$  оцениватель должен иметь вид  $Cy$ , где  $C = C(X)$  — матрица размерностью  $m \times n$ . Мат. ожидание оценки  $Cy$  равно

$$E(Cy) = C E(X\hat{\beta} + \epsilon) = CX\hat{\beta} + E(\epsilon) = CX\hat{\beta}.$$

Ковариационная матрица оценок равна

$$\begin{aligned} \text{Var}(Cy) &= E([Cy - E(Cy)][Cy - E(Cy)]^T) = \\ &= E(C[y - X\hat{\beta}][y - X\hat{\beta}]^T C^T). \end{aligned}$$

Поскольку  $y - X\hat{\beta} = \epsilon$ , то

$$\text{Var}(Cy) = E(C\epsilon\epsilon^T C^T) = CE(\epsilon\epsilon^T)C^T = C(\sigma^2 I)C^T = \sigma^2 CC^T.$$

В случае оценок МНК  $C = (X^T X)^{-1} X^T$ , поэтому

$$\text{Var}(\hat{\beta}_{\text{МНК}}) = \sigma^2 (X^T X)^{-1} X^T ((X^T X)^{-1} X^T)^T = \sigma^2 (X^T X)^{-1}.$$

Несмещенность оценок означает

$$CX\hat{\beta} = \hat{\beta}.$$

Нам требуется, чтобы оцениватель был несмещенным при любом векторе истинных оценок  $\hat{\beta}$ , а это может быть только если  $CX = I$ .

Обозначим

$$D = C - (X^T X)^{-1} X^T.$$

Заметим, что  $DX = CX - (X^T X)^{-1} X^T X = I - I = O$ .

Дисперсию оценок  $Cy$  для произвольной матрицы  $C$ , таким образом, можно переписать в виде

$$\begin{aligned} \text{Var}(Cy) &= \hat{\sigma}^2 C C^T = \hat{\sigma}^2 (D + (X^T X)^{-1} X^T) (D + (X^T X)^{-1} X^T)^T = \\ &= \hat{\sigma}^2 (X^T X)^{-1} + \hat{\sigma}^2 D X (X^T X)^{-1} + \hat{\sigma}^2 (X^T X)^{-1} X^T D^T + \hat{\sigma}^2 D D^T = \\ &= \text{Var}(\hat{\beta}_{\text{МНК}}) + \hat{\sigma}^2 D D^T. \end{aligned}$$

Поскольку матрица  $\hat{\sigma}^2 D D^T$  является положительно полуопределенной, то

$$\text{Var}(Cy) \geq \text{Var}(\hat{\beta}_{\text{МНК}}).$$

Рассмотрим линейную комбинацию оценок с вектором коэффициентов  $\gamma$ :

$$\begin{aligned} \text{Var}(\gamma^T Cy) &= \gamma^T \text{Var}(Cy) \gamma = \gamma^T \text{Var}(\hat{\beta}_{\text{МНК}}) \gamma + \sigma^2 \gamma^T D D^T \gamma = \\ &= \text{Var}(\gamma^T \hat{\beta}_{\text{МНК}}) = \hat{\sigma}^2 \|\gamma^T D\|^2. \end{aligned}$$

Отсюда следует, что

$$\text{Var}(\gamma^T Cy) \geq \text{Var}(\gamma^T \hat{\beta}_{\text{МНК}}).$$

Нам осталось доказать соотношение

$$\text{Var}_Q(Cy) \geq \text{Var}_Q(\hat{\beta}_{\text{МНК}}).$$

Поскольку

$$\text{Var}_Q(\hat{\beta}) = \text{tr}[\text{Var}(\hat{\beta}) Q],$$

то

$$\begin{aligned} \text{Var}_Q(Cy) &= \text{tr}[\text{Var}(Cy) Q] = \text{tr}[(\text{Var}(\hat{\beta}_{\text{МНК}}) + \hat{\sigma}^2 D D^T) Q] = \\ &= \text{tr}[\text{Var}(\hat{\beta}_{\text{МНК}}) Q] + \hat{\sigma}^2 \text{tr}[D D^T Q] = \text{Var}_Q(\hat{\beta}_{\text{МНК}}) + \hat{\sigma}^2 \text{tr}[D^T Q D]. \end{aligned}$$

Пользуясь тем, что по предположению  $Q$  — положительно определенная симметричная матрица, несложно показать, что

$$\hat{\sigma}^2 \text{tr}[D^T Q D] \geq 0. \quad \blacksquare$$

Когда пользуются методом наименьших квадратов, то в качестве оценки дисперсии ошибки обычно берут величину

$$s^2 = \frac{RSS}{n - m}.$$

Эта оценка дисперсии ошибки является несмещенной в предположении, что матрица  $X$  детерминирована. Покажем это. Математическое ожидание суммы квадратов остатков равно

$$E(RSS) = E(e^T e) = E((M\epsilon)^T (M\epsilon)) = E(\epsilon^T M \epsilon) =$$

$$\begin{aligned} &= E(\text{tr}(\boldsymbol{\varepsilon}^T \mathbf{M} \boldsymbol{\varepsilon})) = E(\text{tr}(\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^T \mathbf{M})) = \text{tr}(E(\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^T \mathbf{M})) = \\ &= \hat{\sigma}^2 \text{tr}(\mathbf{I}_n \mathbf{M}) = \hat{\sigma}^2 \text{tr}(\mathbf{M}) = \hat{\sigma}^2 (n - m). \end{aligned}$$

Отсюда

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

Доказывая теорему Гаусса-Маркова, мы вывели, что ковариационная матрица коэффициентов МНК в предположении, что матрица  $\mathbf{X}$  детерминирована, равна

$$\text{Var}(\hat{\boldsymbol{\beta}}_{\text{МНК}}) = \hat{\sigma}^2 (\mathbf{X}^T \mathbf{X})^{-1}.$$

Подставив вместо неизвестной дисперсии ее оценку  $s^2$ , получим оценку матрицы  $\text{Var}(\hat{\boldsymbol{\beta}}_{\text{МНК}})$ :

$$\hat{\text{Var}}(\hat{\boldsymbol{\beta}}_{\text{МНК}}) = s^2 (\mathbf{X}^T \mathbf{X})^{-1}.$$

Очевидно, что эта оценка является несмещенной.

### Асимптотические свойства оценок МНК: состоятельность

Мы рассмотрели выше свойства оценок МНК в конечных выборках. Представляют интерес также асимптотические свойства оценок, т.е. свойства, проявляющиеся при стремлении количества наблюдений к бесконечности. Важнейшим асимптотическим свойством является состоятельность. Оценка параметров называется состоятельной, если она стремится в статистическом смысле к истинному значению параметров по мере того, как растет количество имеющихся наблюдений  $n$ . Обычное определение состоятельности опирается на сходимость по вероятности, т.е. оценка  $\hat{\boldsymbol{\beta}}^{(n)}$  является состоятельной, если

$$\text{plim}_{n \rightarrow \infty} \hat{\boldsymbol{\beta}}^{(n)} = \hat{\boldsymbol{\beta}},$$

что по определению вероятностного предела означает, что для любой положительной константы  $\varepsilon$

$$\lim_{n \rightarrow \infty} \text{Prob}(\|\hat{\boldsymbol{\beta}}^{(n)} - \hat{\boldsymbol{\beta}}\| < \varepsilon) = 1.$$

Идея состоятельности заключается в том, что мы можем потенциально выяснить, каким был истинный вектор  $\hat{\boldsymbol{\beta}}$ , если будем накапливать данные. Состоятельность — основное требование к статистическим оценкам. Вообще говоря, оценку, которая не является состоятельной, обычно не называют

оценкой. Вполне допустимым считается использовать смещенные или неэффективные оценки, но не принято использовать несостоятельные оценки.

Следует, однако, оговориться, что состоятельность — это теоретическое свойство. Исследователь располагает только конкретным конечным набором данных. Ему мало помогает тот факт, что, поскольку используемый им способ оценивания состоятелен, то, если бы он имел бесконечно много данных, то он бы точно знал истинные параметры. С этой точки зрения подход, берущий за основу функцию потерь, кажется более правильным.

При рассмотрении асимптотических свойств тонкий момент состоит в том, каким образом добавляются новые наблюдения к имеющейся выборке. Есть два основных способа:

(1) Матрица регрессоров повторяется. Пусть мы рассматриваем регрессию с  $n = \bar{n}$  и  $X^{(n)} = \bar{X}$ . Тогда при  $n = 2\bar{n}$  матрица регрессоров должна быть равна

$$X^{(2n)} = \begin{bmatrix} \bar{X} \\ \bar{X} \end{bmatrix}.$$

(2) Задано правило, в соответствии с которым порождаются регрессоры. Если  $j$ -й регрессор детерминированный, то его обычно можно задать как функцию номера наблюдения, т.е.

$$x_{ij} = f_j(i).$$

Типичный пример здесь — линейный тренд, когда  $x_{ij} = i$ .

Регрессор может быть также выборкой некоторой случайной величины (с не изменяющимся от номера наблюдения распределением), т.е.  $x_{ij}$  — независимые одинаково распределенные случайные величины.

Можно рассматривать более сложные порождающие процессы, например, авторегрессионный процесс:

$$x_{ij} = \mu + \rho x_{i-1j} + \xi_i,$$

где  $\xi_i$  — независимые одинаково распределенные случайные величины с нулевым мат. ожиданием ( $\xi_i \sim \text{IID}(0, \sigma^2)$ ).

Если регрессоры не независимы между собой или не являются независимыми для разных наблюдений, то нужно рассматривать, вообще говоря, совместное распределение.

Докажем, что при некоторых условиях оценки МНК являются состоятельными. Мы используем при доказательстве следующие правила операций с вероятностными пределами:

- (•) Если пределы  $\text{plim}_{n \rightarrow \infty} \mathbf{A}^{(n)}$  и  $\text{plim}_{n \rightarrow \infty} \mathbf{B}^{(n)}$  существуют, то  $\text{plim}_{n \rightarrow \infty} (\mathbf{A}^{(n)} + \mathbf{B}^{(n)}) = \text{plim}_{n \rightarrow \infty} \mathbf{A}^{(n)} + \text{plim}_{n \rightarrow \infty} \mathbf{B}^{(n)}$ .
- (•) Если пределы  $\text{plim}_{n \rightarrow \infty} \mathbf{A}^{(n)}$  и  $\text{plim}_{n \rightarrow \infty} \mathbf{B}^{(n)}$  существуют, то  $\text{plim}_{n \rightarrow \infty} (\mathbf{A}^{(n)} \mathbf{B}^{(n)}) = \text{plim}_{n \rightarrow \infty} \mathbf{A}^{(n)} \text{plim}_{n \rightarrow \infty} \mathbf{B}^{(n)}$ .
- (•) Если  $\text{plim}_{n \rightarrow \infty} \mathbf{A}^{(n)} = \mathbf{A}^{(\infty)}$  существует, и  $\det(\mathbf{A}^{(\infty)}) \neq 0$ , то  $\text{plim}_{n \rightarrow \infty} ((\mathbf{A}^{(n)})^{-1}) = (\mathbf{A}^{(\infty)})^{-1}$ .

**Теорема** (состоятельность оценок МНК).

Пусть существуют пределы

$$\text{plim}_{n \rightarrow \infty} \left[ \frac{1}{n} \mathbf{X}^{(n)\top} \mathbf{X}^{(n)} \right] = \mathbf{S} \quad \text{и} \quad \text{plim}_{n \rightarrow \infty} \left[ \frac{1}{n} \mathbf{X}^{(n)\top} \boldsymbol{\epsilon}^{(n)} \right],$$

причем

(а)  $\det(\mathbf{S}) \neq 0$ ,

(б)  $\text{plim}_{n \rightarrow \infty} \left[ \frac{1}{n} \mathbf{X}^{(n)\top} \boldsymbol{\epsilon}^{(n)} \right] = \mathbf{0}_m$ .

Тогда  $\text{plim}_{n \rightarrow \infty} \hat{\boldsymbol{\beta}}_{\text{МНК}}^{(n)} = \hat{\boldsymbol{\beta}}$ .

Замечание:

Условие (а) — это условие асимптотической идентифицируемости. Условие (б) — это условие асимптотической ортогональности регрессоров и ошибки.

Доказательство:

(Мы будем опускать индекс  $n$  для упрощения записи.)

Поскольку  $\hat{\boldsymbol{\beta}}_{\text{МНК}} = \hat{\boldsymbol{\beta}} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\epsilon}$ , то

$$\begin{aligned} \text{plim} \hat{\boldsymbol{\beta}}_{\text{МНК}} &= \hat{\boldsymbol{\beta}} + \text{plim} \left[ \left( \frac{1}{n} \mathbf{X}^\top \mathbf{X} \right)^{-1} \left( \frac{1}{n} \mathbf{X}^\top \boldsymbol{\epsilon} \right) \right] = \\ &= \hat{\boldsymbol{\beta}} + \text{plim} \left[ \left( \frac{1}{n} \mathbf{X}^\top \mathbf{X} \right)^{-1} \right] \text{plim} \left[ \frac{1}{n} \mathbf{X}^\top \boldsymbol{\epsilon} \right] = \\ &= \hat{\boldsymbol{\beta}} + \text{plim} \left[ \frac{1}{n} \mathbf{X}^\top \mathbf{X} \right]^{-1} \text{plim} \left[ \frac{1}{n} \mathbf{X}^\top \boldsymbol{\epsilon} \right] = \\ &= \hat{\boldsymbol{\beta}} + \mathbf{S}^{-1} \mathbf{0}_m = \hat{\boldsymbol{\beta}}. \quad \blacksquare \end{aligned}$$

Теорема требует некоторых комментариев.

Откуда в условии теоремы множитель  $\frac{1}{n}$ ?

Если матрица регрессоров детерминированная и получается повторением некоторой исходной матрицы  $\bar{\mathbf{X}}$ , т.е.

$$\mathbf{X}^{(ka)} = \begin{bmatrix} \bar{\mathbf{X}} \\ \vdots \\ \bar{\mathbf{X}} \end{bmatrix} \Bigg\} k ,$$

то

$$\lim_{n \rightarrow \infty} \left[ \frac{1}{n} \mathbf{X}^{(n)\top} \mathbf{X}^{(n)} \right] = \bar{\mathbf{X}}^\top \bar{\mathbf{X}}.$$

(В данном случае вероятностный предел есть просто обычный предел.)

Если матрица регрессоров состоит из независимых одинаково распределенных строк, то

$$\text{plim}_{n \rightarrow \infty} \left[ \frac{1}{n} \mathbf{X}^{(n)\top} \mathbf{X}^{(n)} \right] = \mathbf{E}(\mathbf{X}_i^\top \mathbf{X}_i)$$

— это закон больших чисел для векторных случайных величин (он выполнен, если верны некоторые условия регулярности).

Рассмотрим также условие  $\text{plim}_{n \rightarrow \infty} \left[ \frac{1}{n} \mathbf{X}^{(n)\top} \boldsymbol{\varepsilon}^{(n)} \right] = \mathbf{0}_m$ . Если матрица  $[\mathbf{X}^{(n)}, \boldsymbol{\varepsilon}^{(n)}]$  (матрица регрессоров, дополненная вектором ошибок) состоит из независимых одинаково распределенных строк, то по закону больших чисел

$$\text{plim}_{n \rightarrow \infty} \left[ \frac{1}{n} \mathbf{X}^{(n)\top} \boldsymbol{\varepsilon}^{(n)} \right] = \mathbf{E}(\mathbf{X}_i^\top \boldsymbol{\varepsilon}_i).$$

Если мы предполагаем, как обычно, что ошибки имеют нулевое мат. ожидание ( $\mathbf{E}(\boldsymbol{\varepsilon}_i) = \mathbf{0}$ ), то  $\mathbf{E}(\mathbf{X}_i^\top \boldsymbol{\varepsilon}_i)$  есть ковариация  $\mathbf{X}_i$  и  $\boldsymbol{\varepsilon}_i$ . Таким образом, в рассматриваемом случае (наблюдения независимы и одинаково распределены) условие  $\text{plim}_{n \rightarrow \infty} \left[ \frac{1}{n} \mathbf{X}^{(n)\top} \boldsymbol{\varepsilon}^{(n)} \right] = \mathbf{0}_m$  эквивалентно некоррелированности регрессоров и ошибки.

Приведем примеры вырожденности предела

$$\mathbf{S} = \text{plim}_{n \rightarrow \infty} \left[ \frac{1}{n} \mathbf{X}^{(n)\top} \mathbf{X}^{(n)} \right],$$

в случае чего имеет место асимптотическая неидентифицируемость, и оценки МНК несостоятельны.

Можно рассмотреть два типичных случая.

1) «Затухающий» регрессор. В детерминированном случае это, например, затухающая экспонента:

$$x_{ij} = e^{a-\beta i} \quad (\beta > 0).$$

Случайный регрессор тоже может «затухать»; например,

$$x_{ij} = \xi_i e^{a-\beta i} \quad (\beta > 0),$$

где через  $\xi_i$  обозначены независимые одинаково распределенные случайные величины.

«Затухание» регрессора означает, что асимптотически он неотличим от вектора, состоящего из нулей. Но если один из регрессоров — нулевой вектор, то матрица регрессоров имеет неполный ранг. Таким образом, если регрессор «затухает», то с ростом количества наблюдений матрица регрессоров в определенном смысле все более приближается к матрице неполного ранга.

2) Полная коррелированность регрессоров. Пусть опять матрица регрессоров состоит из независимых одинаково распределенных строк. Тогда, как уже говорилось,  $S = E(X_i^T X_i)$ . Если среди регрессоров есть пара регрессоров ( $j$  и  $k$ ), таких что

$$E(x_{ij}^T x_{ik}) = E(x_{ij}^2) = E(x_{ik}^2),$$

то  $j$ -я и  $k$ -я строки матрицы  $S$  будут совпадать между собой, и, следовательно, матрица  $S$  будет вырожденной.

На самом деле для того, чтобы  $S$  была вырожденной, достаточно, чтобы  $\text{Cov}(x_{ij}, x_{ik}) = 1$  или  $-1$  и среди остальных регрессоров была константа.

Изложенная теорема использует слишком жесткие условия. Есть случаи, когда предел  $\text{plim}_{n \rightarrow \infty} [\frac{1}{n} X^{(n)T} X^{(n)}]$  не существует, а оценки МНК состоятельны. На самом деле в этом случае оценки МНК могут быть сверхсостоятельными, то есть такими, что скорость сходимости к истинным параметрам у них превышает обычную (более формальное определение сверхсостоятельности будет дано ниже??).

Упомянем два характерных случая.

Если один из регрессоров — линейный тренд ( $x_{ij} = i$ ), то

$$\frac{1}{n} x_j^T x_j = \frac{1}{n} \frac{n(n+1)}{2} = \frac{n+1}{2} \rightarrow \infty.$$

Ясно, что в этом случае  $jj$ -й диагональный элемент матрицы  $\frac{1}{n} X^T X$  должен стремиться к бесконечности с ростом  $n$ .

Другой характерный случай — это случай, когда один из регрессоров представляет собой случайное блуждание:

$$x_{ij} = x_{i-1j} + \xi_i,$$

где  $\xi_i \sim \text{IID}(0, \sigma^2)$ .

В этом случае

$$E\left[\frac{1}{n}x_{ij}^T x_{ij} \mid x_{1j}\right] = x_{1j}^2 + n\sigma^2.$$

(Берем условное мат. ожидание, поскольку безусловного здесь просто не существует.) Поэтому

$$\text{plim}\left[\frac{1}{n}x_j^T x_j\right] = \infty.$$

Понятно, что если в регрессии есть такого рода регрессоры, то дополнительные наблюдения дают много новой информации о коэффициентах при этих регрессорах — в этом причина сверхсостоятельности. (С другой стороны, наличие регрессоров такого рода может быть опасным, поскольку в этом случае можно столкнуться с проблемой ложной регрессии. См.???)

Можно рассмотреть также асимптотические свойства оценки дисперсии ошибки

$$s^2 = \frac{RSS}{n-m}.$$

**Теорема** (состоятельность оценки  $s^2$ ).

Пусть выполнены следующие условия

⊕ ошибки независимы и одинаково распределены, причем  $E[\varepsilon_i] = 0$ ,  $E[\varepsilon_i^2] = \sigma^2$ ;

⊕  $\hat{\beta}$  — состоятельная оценка истинного вектора параметров  $\beta$ ;

⊕ существуют пределы

$$\text{plim}\left[\frac{1}{n}X^T X\right] = S \quad \text{и} \quad \text{plim}\left[\frac{1}{n}X^T \varepsilon\right] = s.$$

Тогда  $\text{plim}\left[\frac{1}{n}RSS\right] = \sigma^2$ .

Замечания:

1) Поскольку  $\text{plim}\left[\frac{1}{n}RSS\right] = \sigma^2$ , то  $\text{plim}\left[\frac{1}{n-m}RSS\right] = \sigma^2$  (в пределе множители  $\frac{1}{n}$  и  $\frac{1}{n-m}$  неотличимы). Значит, из данной теоремы следует состоятельность  $s^2 = \frac{1}{n-m}RSS$  как оценки дисперсии.

2) В теореме не используется то, что  $\hat{\beta}$  — оценки МНК. Она верна для любого состоятельного вектора оценок.

Доказательство:

Выразим остатки через ошибки следующим образом

$$e = y - X\hat{\beta} = y - X\hat{\beta} + X(\hat{\beta} - \beta) = \epsilon + X(\hat{\beta} - \beta).$$

Отсюда

$$\begin{aligned} \text{plim}\left[\frac{RSS}{n}\right] &= \text{plim}\left[\frac{1}{n}e^T e\right] = \text{plim}\left[\frac{1}{n}\epsilon^T \epsilon\right] + \text{plim}\left[\left(\frac{1}{n}X^T \epsilon\right)^T (\hat{\beta} - \beta)\right] + \\ &+ \text{plim}\left[(\hat{\beta} - \beta)^T \left(\frac{1}{n}X^T \epsilon\right)\right] + \text{plim}\left[(\hat{\beta} - \beta)^T \left(\frac{1}{n}X^T X\right)(\hat{\beta} - \beta)\right] = \\ &= \text{plim}\left[\frac{1}{n}\epsilon^T \epsilon\right] + \text{plim}\left[\frac{1}{n}X^T \epsilon\right]^T \text{plim}[\hat{\beta} - \beta] + \text{plim}[\hat{\beta} - \beta]^T \text{plim}\left[\frac{1}{n}X^T \epsilon\right] + \\ &+ \text{plim}[\hat{\beta} - \beta]^T \text{plim}\left[\frac{1}{n}X^T X\right] \text{plim}[\hat{\beta} - \beta] = \\ &= \text{plim}\left[\frac{1}{n}\epsilon^T \epsilon\right] + s^T 0 + 0^T s + 0^T S 0 = \text{plim}\left[\frac{1}{n}\epsilon^T \epsilon\right]. \end{aligned}$$

Здесь мы воспользовались тем, что, поскольку  $\hat{\beta}$  — состоятельная оценка, то

$$\text{plim}[\hat{\beta} - \beta] = 0.$$

Менее формально это можно доказать так. Поскольку  $\hat{\beta}$  приближается к  $\beta$  с ростом количества наблюдений, то  $e = y - X\hat{\beta}$  приближается к  $\epsilon = y - X\beta$ . (В каком-то смысле, остатки — состоятельные оценки ошибок.) Следовательно,  $\frac{1}{n}e^T e$  приближается к  $\frac{1}{n}\epsilon^T \epsilon$ .

Поскольку ошибки независимы, то их квадраты тоже независимы. Кроме того, мат. ожидание квадратов ошибок равно  $\hat{\sigma}^2$ . Таким образом, по закону больших чисел

$$\text{plim}\left[\frac{1}{n}\epsilon^T \epsilon\right] = \text{plim}\left[\frac{1}{n}\sum_{i=1}^n \epsilon_i^2\right] = \hat{\sigma}^2.$$

Тем самым, доказано, что  $\text{plim}\left[\frac{RSS}{n}\right] = \hat{\sigma}^2$ .

■

Если матрица регрессоров детерминирована, то, как мы показали выше,

$$\text{Var}(\hat{\beta}) = \hat{\sigma}^2 (X^T X)^{-1}.$$

В общем случае это равенство выполнено только условно по  $X$ :

$$\text{Var}(\hat{\beta}|X) = \hat{\sigma}^2 (X^T X)^{-1}.$$

Взяв математические ожидания от обеих частей, получим

$$\text{Var}(\hat{\beta}) = E(\text{Var}(\hat{\beta}|X)) = \hat{\sigma}^2 E((X^T X)^{-1}).$$

Величину

$$\hat{\text{Var}}(\hat{\beta}) = s^2 (X^T X)^{-1}$$

при выполнении ряда предположений можно использовать в качестве оценки  $\text{Var}(\hat{\beta})$  в случае случайной матрицы регрессоров, так как она будет состоятельной оценкой.

Пусть, например, строки матрицы регрессоров независимы и одинаково распределены. По закону больших чисел

$$\text{plim}\left[\frac{1}{n} X^T X\right] = E(X_i^T X_i) = \frac{1}{n} E(X^T X).$$

Кроме того, как только что доказано,  $\text{plim}[s^2] = \hat{\sigma}^2$ . Поэтому

$$\begin{aligned} \text{plim}[n \hat{\text{Var}}(\hat{\beta})] &= \text{plim}[s^2 n (X^T X)^{-1}] = \\ &= \text{plim}[s^2] \text{plim}\left[\frac{1}{n} X^T X\right]^{-1} = \hat{\sigma}^2 \left(\frac{1}{n} E(X^T X)\right)^{-1} = n \hat{\sigma}^2 (E(X^T X))^{-1}. \end{aligned}$$

(?????????????)

### Свойства функций от коэффициентов МНК

Пусть нам требуется получить оценки параметров  $\gamma \in \mathbb{R}^k$ , которые являются функциями исходных коэффициентов:

$$\gamma = g(\beta).$$

Естественно использовать оценки следующего вида:

$$\hat{\gamma} = g(\hat{\beta}).$$

Если  $g(\cdot)$  — непрерывная функция, то, поскольку  $\hat{\beta}$  — состоятельные оценки  $\beta$ , то по теореме Слущкого  $\hat{\gamma}$  должны быть состоятельными оценками величины  $\gamma = g(\beta)$ .

Будут ли эти оценки несмещенными? В общем случае нет. Однако если функция  $g(\cdot)$  линейна, то (при стандартных предположениях)  $\hat{\gamma}$  несмещенно оценивает  $\gamma$ .

Проведем неформальный вывод оценки матрицы ковариаций оценок  $\hat{\gamma}$ . Пусть функция  $g(\cdot)$  имеет непрерывную производную. Тогда можно разложить ее в ряд Тейлора в окрестности точки  $\beta$ .

$$g(\hat{\beta}) \approx g(\beta) + \frac{\partial g(\beta)}{\partial \beta} (\hat{\beta} - \beta) = \gamma + G(\beta) (\hat{\beta} - \beta),$$

или

$$\hat{\gamma} - \gamma \approx G(\hat{\beta}) (\hat{\beta} - \beta),$$

где

$$G(\beta) = \frac{\partial g(\beta)}{\partial \beta}$$

— матрица Якоби для функции  $g(\cdot)$ .

Отсюда можно приближенно записать

$$\text{Var}(\hat{\gamma}) \approx G(\hat{\beta}) \text{Var}(\hat{\beta}) G(\hat{\beta})^T.$$

Если функция  $g(\cdot)$  линейна, то равенство здесь точное.

Если в вместо матрицы  $\text{Var}(\hat{\beta})$  подставим ее состоятельную оценку, то получим состоятельную оценку матрицы  $\text{Var}(\hat{\gamma})$ . Например, если взять

$$\hat{\text{Var}}(\hat{\beta}) = s^2 (X^T X)^{-1},$$

то имеем

$$\hat{\text{Var}}(\hat{\gamma}) = s^2 G(\hat{\beta}) (X^T X)^{-1} G(\hat{\beta})^T.$$

В частном случае линейной функции  $g(\cdot)$ ...

### Асимптотические свойства оценок МНК: сходимость в среднеквадратическом

Мы рассмотрели состоятельность в смысле сходимости по вероятности. Точность оценок (а следовательно, и сходимость) можно рассматривать также с точки зрения функции ожидаемых потерь. Сходимость по вероятности тоже можно рассматривать с точки зрения функции ожидаемых потерь.

Пусть

$$L(\beta, \hat{\beta}) = \begin{cases} 0, & \text{если } \|\beta - \hat{\beta}\| < \Delta \\ 1, & \text{если } \|\beta - \hat{\beta}\| > \Delta \end{cases}.$$

Тогда  $E(L(\beta, \hat{\beta})) = 1 - \text{Prob}(\|\hat{\beta} - \beta\| < \Delta)$ . Состоятельность оценки  $\hat{\beta}$  означает, что ожидаемые потери стремятся к нулю ( $E(L(\beta, \hat{\beta})) \rightarrow 0$ ), какую бы границу  $\Delta$  мы не взяли в функции потерь.

Более интересна, однако, квадратичная функция потерь

$$L(\beta, \hat{\beta}) = (\beta - \hat{\beta})^T Q (\beta - \hat{\beta}).$$

Мы хотим, чтобы ожидаемые потери стремились к нулю с ростом количества имеющихся наблюдений. Оказывается,

это эквивалентно состоятельности в смысле среднеквадратичной сходимости, т.е.

$$MSE_Q(\hat{\beta}) \xrightarrow{ms} 0 \quad \Leftrightarrow \quad \hat{\beta} \xrightarrow{ms} \beta,$$

где символ  $\xrightarrow{ms}$  обозначает сходимость в среднеквадратическом.

Хотя мы не можем (как объяснялось выше) минимизировать среднеквадратическую ошибку не обладая априорной информацией, но при определенных условиях среднеквадратическая ошибка оценок МНК стремится к нулю с ростом количества наблюдений.

Из сходимости в среднеквадратическом следует сходимости по вероятности, поэтому

$$MSE_Q(\hat{\beta}) \rightarrow 0 \Rightarrow \text{plim } \hat{\beta} = \beta.$$

**Теорема** (состоятельность оценок МНК в смысле среднеквадратической сходимости).

Пусть  $\lambda_1, \dots, \lambda_m$  — собственные числа матрицы  $X^T X$ .

Пусть

(а) матрица регрессоров  $X$  детерминирована,

(б)  $E(\epsilon) = 0$ ,

(в)  $E(\epsilon\epsilon^T) = \sigma^2 I$ .

Тогда

$$MSE_Q(\hat{\beta}_{\text{МНК}}) \rightarrow 0$$

для произвольной положительно определенной матрицы  $Q$  (подходящей размерности) выполнено тогда и только тогда, когда

$$\lambda_j \rightarrow \infty \quad \forall j.$$

**Доказательство:**

Если условия теоремы верны, то оценки МНК несмещенные и

$$MSE_Q(\hat{\beta}_{\text{МНК}}) = \text{tr}[\text{Var}(\hat{\beta}_{\text{МНК}}) Q].$$

Условия теоремы также гарантируют, что

$$\text{Var}(\hat{\beta}_{\text{МНК}}) = \sigma^2 (X^T X)^{-1}$$

(см. теорему Гаусса-Маркова).

Поэтому

$$MSE_Q(\hat{\beta}_{\text{МНК}}) = \sigma^2 \text{tr}[(X^T X)^{-1} Q].$$

Поскольку матрица  $X^T X$  положительно определена, то существует следующее ее представление (так называемое спектральное разложение):

$$X^T X = H^T \Lambda H,$$

где  $H$  — ортогональная матрица (составленная из собственных векторов матрицы  $X^T X$ ),  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_m)$ .

С помощью этого разложения легко найти обратную матрицу:

$$(X^T X)^{-1} = H^T \Lambda^{-1} H.$$

Таким образом,

$$\begin{aligned} \text{MSE}_Q(\hat{\beta}_{\text{МНК}}) &= \sigma^2 \text{tr}[(X^T X)^{-1} Q] = \sigma^2 \text{tr}(H^T \Lambda^{-1} H Q) = \\ &= \sigma^2 \text{tr}(H Q H^T \Lambda^{-1}) = \sigma^2 \text{tr}(\Lambda \Lambda^{-1}). \end{aligned}$$

Здесь мы воспользовались правилом  $\text{tr}(BC) = \text{tr}(CB)$  и обозначили  $A = H Q H^T$ .

Пусть  $a_{jj}$  —  $j$ -й диагональный элемент матрицы  $A$ . Все  $a_{jj}$  положительны. Это можно показать следующим образом:

$$a_{jj} = e^{jT} A e^j = e^{jT} H Q H^T e^j = h_j^T Q h_j > 0.$$

Здесь  $h_j^T$  —  $j$ -я строка матрицы  $H$ . Последнее неравенство следует из того, что  $Q$  — положительно определенная матрица и  $h_j \neq 0$ .

Мы получили следующее представление среднеквадратической ошибки:

$$\text{MSE}_Q(\hat{\beta}_{\text{МНК}}) = \sigma^2 \text{tr}(\Lambda \Lambda^{-1}) = \sigma^2 \left( \frac{a_{11}}{\lambda_1} + \dots + \frac{a_{mm}}{\lambda_m} \right).$$

Поскольку матрица  $X^T X$  — положительно определенная, то все  $\lambda_j$  положительны. Отсюда следует доказываемое утверждение:

Среднеквадратическая ошибка стремиться к нулю тогда и только тогда, когда каждое из собственных чисел матрицы  $X^T X$  неограниченно возрастает. ■

Если матрица регрессоров детерминированная и получается повторением некоторой исходной матрицы  $\bar{X}$ , т.е.

$$X^{(k\bar{n})} = \begin{bmatrix} \bar{X} \\ \vdots \\ \bar{X} \end{bmatrix} \Bigg\} k,$$

то

$$X^{(k\bar{n})T} X^{(k\bar{n})} = k \bar{X}^T \bar{X}$$

и

$$\lambda^{(k\bar{n})} = k\lambda_j^{(\bar{n})}.$$

Отсюда следует, что  $\text{MSE}_Q(\hat{\beta}_{\text{МНК}}^{(k\bar{n})}) \rightarrow 0$  при  $k \rightarrow \infty$ .

### Следствия нормальности ошибок

Нормальность ошибок мы будем рассматривать только в случае, когда матрица регрессоров  $X$  детерминирована. Также будем предполагать, что ошибки имеют нулевое мат. ожидание  $E(\epsilon) = 0$  и что  $E(\epsilon\epsilon^T) = \sigma^2 I$ . (Для краткости, когда будем говорить о нормальности, будем подразумевать, что эти два условия тоже выполнены.)

Таким образом, предположим дополнительно, что ошибки имеют нормальное распределение (гауссовские ошибки). При  $E(\epsilon) = 0$  и  $E(\epsilon\epsilon^T) = \sigma^2 I$ , ошибки некоррелированы ( $\text{Var}(\epsilon) = \sigma^2 I$ ), и, тем самым, независимы. (По свойствам нормального распределения, если нормально распределенные величины некоррелированы, то они независимы.) Каждая ошибка имеет нулевое мат. ожидание и дисперсию  $\sigma^2$ .

Мы можем записать кратко, что

$$\epsilon_i \sim \mathcal{NID}(0, \sigma^2)$$

— т.е. ошибки независимы и имеют одно и то же нормальное распределение с параметрами 0 и  $\sigma^2$ .

Вектор  $\epsilon$  имеет многомерное нормальное распределение с параметрами 0 и  $\sigma^2 I$ :

$$\epsilon \sim \mathcal{N}(0_n, \sigma^2 I_n).$$

Плотность многомерного нормального распределения с параметрами  $a$  (мат. ожидание) и  $A$  (матрица ковариаций) имеет вид:

$$p(x) = (2\pi)^{-\frac{n}{2}} |A|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-a)^T A^{-1}(x-a)}.$$

По этой формуле находим, подставляя  $a = 0_n$  и  $A = \sigma^2 I_n$ , что плотность распределения ошибок равна

$$p_\epsilon(x) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2} x^T x}.$$

Предположение  $E(\epsilon\epsilon^T) = \sigma^2 I$  часто называют сферичностью ошибок. Действительно, точки, в которых плотность одинакова, составляют сферу в  $\mathbb{R}^n$ , задаваемую уравнением  $x^T x = \|x\|^2 = \text{const}$ . Если бы ковариационная матрица ошибок  $E(\epsilon\epsilon^T)$  не была диагональной, либо же была диагональной, но с раз-

ными диагональными элементами, то рассматриваемые точки составляли бы уже не  $n$ -мерную сферу, а  $n$ -мерный эллипсоид.

В предположении нормальности оценка МНК является оценкой максимального правдоподобия (ММП).

Функция правдоподобия — это для непрерывных распределений плотность вероятности. Зависимая переменная распределена как

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \hat{\sigma}^2 \mathbf{I}_n).$$

Здесь мы воспользовались следующим свойством нормального распределения:

Если  $\mathbf{x} \sim \mathcal{N}(\mathbf{a}, \mathbf{A})$ , то  $\mathbf{b} + \mathbf{Bx} \sim \mathcal{N}(\mathbf{b} + \mathbf{Ba}, \mathbf{BAB}^\top)$ .  
(Этим свойством мы будем часто пользоваться в дальнейшем.)

Таким образом, функция правдоподобия для линейной регрессии равна

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right).$$

Удобнее использовать логарифмическую функцию правдоподобия:

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} RSS(\boldsymbol{\beta}),$$

где  $RSS(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$

Найдем оценки ММП из условий первого порядка, приравняв градиент нулю:

$$\begin{aligned} \frac{\partial \ell}{\partial \boldsymbol{\beta}} &= -\frac{1}{2\sigma^2} \frac{d RSS(\boldsymbol{\beta})}{d \boldsymbol{\beta}} = \frac{1}{\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{X} = 0, \\ \frac{\partial \ell}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} - \frac{1}{2\sigma^4} RSS(\boldsymbol{\beta}) = 0. \end{aligned}$$

Отсюда находим

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{ММП}} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}, \\ \hat{\sigma}^2_{\text{ММП}} &= \frac{1}{n} RSS(\hat{\boldsymbol{\beta}}_{\text{ММП}}) = \frac{RSS}{n}. \end{aligned}$$

Оценка ММП коэффициентов  $\boldsymbol{\beta}$  совпадает с оценкой МНК. Оценка дисперсии является смещенной. Действительно, поскольку, как мы видели раньше,  $E(RSS) = \hat{\sigma}^2(n - m)$ ,

то

$$E(\hat{\sigma}^2_{\text{ММП}}) = \hat{\sigma}^2 \frac{n - m}{n}.$$

Как известно из статистической теории, оценки ММП (при определенных условиях) являются асимптотически эффективными. Таким образом, оценки МНК в предположении нормальности ошибок являются асимптотически эффективными (имеют наименьшую асимптотическую дисперсию среди состоятельных оценок).

Найдем, исходя из нормальности, распределения различных величин, относящихся к МНК. Выразим через вектор ошибок  $\mathbf{e}$  величины  $\hat{\boldsymbol{\beta}}$ ,  $\hat{\mathbf{y}}$  и  $\mathbf{e}$ :

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \mathring{\boldsymbol{\beta}} + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{e}, \\ \hat{\mathbf{y}} &= \mathbf{X} \mathring{\boldsymbol{\beta}} + \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{e} = \mathbf{X} \mathring{\boldsymbol{\beta}} + \mathbf{P} \mathbf{e}, \\ \mathbf{e} &= (\mathbf{I}_n - \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) \mathbf{e} = \mathbf{M} \mathbf{e}.\end{aligned}$$

Напомним, что если

$$\mathbf{x} \sim \mathcal{N}(\mathbf{a}, \mathbf{A}),$$

то

$$\mathbf{b} + \mathbf{B}\mathbf{x} \sim \mathcal{N}(\mathbf{b} + \mathbf{B}\mathbf{a}, \mathbf{B}\mathbf{A}\mathbf{B}^T).$$

Поскольку  $\mathbf{e} \sim \mathcal{N}(\mathbf{0}_n, \mathring{\sigma}^2 \mathbf{I}_n)$ , то

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\mathring{\boldsymbol{\beta}}, \mathring{\sigma}^2 (\mathbf{X}^T \mathbf{X})^{-1})$$

(учитывая, что  $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{I}_n \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} = (\mathbf{X}^T \mathbf{X})^{-1}$ ),

$$\hat{\mathbf{y}} \sim \mathcal{N}(\mathbf{X} \mathring{\boldsymbol{\beta}}, \mathring{\sigma}^2 \mathbf{P})$$

(учитывая, что  $\mathbf{P}\mathbf{P}^T = \mathbf{P}$ ),

$$\mathbf{e} \sim \mathcal{N}(\mathbf{0}_n, \mathring{\sigma}^2 \mathbf{M})$$

(учитывая, что  $\mathbf{M}\mathbf{M}^T = \mathbf{M}$ ).

Заметим, что векторы  $\hat{\mathbf{y}}$  и  $\mathbf{e}$  некоррелированы:

$$\begin{aligned}\mathbb{E}((\hat{\mathbf{y}} - \mathbf{X} \mathring{\boldsymbol{\beta}}) \mathbf{e}^T) &= \mathbb{E}(\mathbf{P} \mathbf{e} (\mathbf{M} \mathbf{e})^T) = \\ &= \mathbf{P} \mathbb{E}(\mathbf{e} \mathbf{e}^T) \mathbf{M} = \mathring{\sigma}^2 \mathbf{P} \mathbf{M} = \mathring{\sigma}^2 \mathbf{0} = \mathbf{0}.\end{aligned}$$

Поскольку  $\hat{\mathbf{y}}$  и  $\mathbf{e}$  имеют нормальное распределение, то  $\hat{\mathbf{y}}$  и  $\mathbf{e}$  независимы. Точно также независимы  $\hat{\boldsymbol{\beta}}$  и  $\mathbf{e}$ .

Имеет место также следующее свойство нормального распределения:

Если  $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ , и  $\mathbf{A}$  — симметричная идемпотентная матрица, то

$$\mathbf{x}^T \mathbf{A} \mathbf{x} \sim \chi^2(k)$$

— величина  $\mathbf{x}^T \mathbf{A} \mathbf{x}$  распределена как хи-квадрат с  $k$  степенями свободы, где  $k = \text{rank}(\mathbf{A})$  ( $= \text{tr}(\mathbf{A})$ ).

Поскольку  $\mathbf{P}$  и  $\mathbf{M}$  — симметричные идемпотентные матрицы, имеющие ранги  $m$  и  $n - m$  соответственно, то

$$\frac{1}{\hat{\sigma}^2}(\hat{y} - X\hat{\beta})^\top(\hat{y} - X\hat{\beta}) = \frac{1}{\hat{\sigma}^2}(\hat{\beta} - \hat{\beta})^\top(X^\top X)^{-1}(\hat{\beta} - \hat{\beta}) \sim \chi^2(m),$$

и

$$\frac{1}{\hat{\sigma}^2}e^\top e = \frac{RSS}{\hat{\sigma}^2} \sim \chi^2(n-m).$$

Поскольку  $\hat{y}$  и  $e$  независимы, то эти две величины тоже независимы.

## Линейные ограничения и проверка гипотез в регрессии

### Оценки МНК при линейных ограничениях

Предположим, что мы ищем только такие оценки МНК  $\beta$ , которые удовлетворяют системе линейных ограничений, которые можно записать в матричном виде:

$$R\beta = r,$$

где  $R$  — матрица  $k \times m$  ( $k < m$ ), имеющая полный ранг,  $r$  — вектор-столбец длиной  $k$ .

Минимизировать сумму квадратов остатков  $RSS(\beta) = (y - X\beta)^\top(y - X\beta)$  при данных ограничениях можно используя теорему Лагранжа. Лагранжиан этой задачи равен

$$\mathcal{L}(\beta, \lambda) = (y - X\beta)^\top(y - X\beta) + \lambda^\top(R\beta - r),$$

где  $\lambda$  — вектор множителей Лагранжа.

$$\frac{\partial \mathcal{L}}{\partial \beta} = -2y^\top X + 2\beta^\top X^\top X + \lambda^\top R,$$

Приравняв данную производную нулю, получим

$$\hat{\beta}_R = (X^\top X)^{-1}X^\top y - \frac{1}{2}(X^\top X)^{-1}R^\top \lambda = \hat{\beta} - \frac{1}{2}(X^\top X)^{-1}R^\top \lambda,$$

Поскольку оценки  $\hat{\beta}_R$  должны удовлетворять ограничениям, то

$$r = R\hat{\beta}_R = R\hat{\beta} - \frac{1}{2}R(X^\top X)^{-1}R^\top \lambda.$$

Выразим отсюда  $\lambda$ :

$$\lambda = 2(R(X^\top X)^{-1}R^\top)^{-1}(R\hat{\beta} - r).$$

Окончательно получаем

$$\hat{\beta}_R = \hat{\beta} - (X^\top X)^{-1}R^\top(R(X^\top X)^{-1}R^\top)^{-1}(R\hat{\beta} - r),$$

Эта формула позволяет получать оценки в регрессии с ограничениями, зная оценки в регрессии без ограничений.

Часто более удобным бывает другой подход, который позволяет получить оценки  $\beta_r$  в результате оценивания модифицированной регрессии без ограничений. Идея этого метода состоит в том, чтобы уменьшить количество оцениваемых коэффициентов, и тогда это меньшее количество параметров можно будет однозначно оценить. Если в исходной регрессии было  $m$  регрессоров и на коэффициенты наложено  $k$  линейно независимых ограничений, то в модифицированной регрессии останется  $m - k$  неизвестных параметров.

Рассмотрим этот метод в простейшем случае. Пусть на последние  $k$  регрессоров наложены «нулевые» ограничения:

$$\beta_j = 0, \quad j = m - k + 1, \dots, m.$$

Тогда, как несложно понять, в модифицированной регрессии нужно оставить первые  $m - k$  регрессоров. Если  $X^*$  — матрица, составленная из этих регрессоров, то оценка в модифицированной регрессии равна

$$\hat{\beta}^* = (X^{*\top} X^*)^{-1} X^{*\top} y$$

а оценка МНК с ограничениями в этих обозначениях равна

$$\hat{\beta}_r = \begin{pmatrix} \hat{\beta}^* \\ 0_k \end{pmatrix}.$$

Это тривиальный случай. В более общем случае бывает возможно используя ограничения  $R\beta = r$  выразить  $k$  коэффициентов через остальные  $m - k$  коэффициентов. Без потери общности можно считать, что упомянутые  $m - k$  коэффициентов относятся к первым  $m - k$  столбцам матрицы  $X$ . Ограничения тогда можно записать в блочном виде:

$$\begin{bmatrix} R_1 & R_2 \end{bmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = r,$$

или

$$R_1 \beta_1 + R_2 \beta_2 = r.$$

Если матрица  $R_2$  (она квадратная  $k \times k$ ) невырождена, то можем выразить  $\beta_2$  через  $\beta_1$ :

$$\beta_2 = R_2^{-1} (r - R_1 \beta_1).$$

Подставим  $\beta_2$  в исходную регрессию, которую можно переписать в виде

$$y = X_1 \beta_1 + X_2 \beta_2 + \varepsilon.$$

После преобразований получим

$$y - X_2 R_2^{-1} r = (X_1 - X_2 R_2^{-1} R_1) \beta_1 + \epsilon.$$

Таким образом, в преобразованной регрессии используются следующие переменные:

$$\begin{aligned} y^* &= y - X_2 R_2^{-1} r, \\ X^* &= X_1 - X_2 R_2^{-1} R_1. \end{aligned}$$

Пусть  $\hat{\beta}^* = (X^{*\top} X^*)^{-1} X^{*\top} y^*$  — оценки в модифицированной регрессии. Это первая часть вектора оценок  $\hat{\beta}_r$  (соответствующая  $\beta_1$ ). Оставшуюся часть вектора  $\hat{\beta}_r$  (соответствующую  $\beta_2$ ) находим с помощью полученного ранее представления  $\beta_2$  через  $\beta_1$ . Т.е.

$$\hat{\beta}_r = \begin{pmatrix} \hat{\beta}^* \\ R_2^{-1}(r - R_1 \hat{\beta}^*) \end{pmatrix}.$$

Опишем также несколько более абстрактный метод. Пусть матрица  $A$  размерностью  $(m-k) \times m$  такова, что (квадратная  $m \times m$ ) матрица  $\begin{bmatrix} A \\ R \end{bmatrix}$  невырождена. Матрицу  $A$  можно взять в качестве матрицы преобразования коэффициентов исходной регрессии в коэффициенты модифицированной:

$$\beta^* = A\beta.$$

Тогда, учитывая ограничения, можно выразить  $\beta$  через  $\beta^*$ . Поскольку

$$\begin{bmatrix} A \\ R \end{bmatrix} \beta = \begin{pmatrix} \beta^* \\ r \end{pmatrix},$$

то

$$\beta = \begin{bmatrix} A \\ R \end{bmatrix}^{-1} \begin{pmatrix} \beta^* \\ r \end{pmatrix}.$$

Пусть

$$\begin{bmatrix} A \\ R \end{bmatrix}^{-1} = \begin{bmatrix} B & C \end{bmatrix},$$

где блок  $B$  соответствует первым  $m-k$  столбцам, а  $C$  — последним  $k$  столбцам. Тогда

$$\beta = B\beta^* + Cr.$$

Подставив это выражение в исходную регрессию, —

$$y = X\beta + \epsilon = X(B\beta^* + Cr) + \epsilon, —$$

получим переменные модифицированной регрессии:

$$\begin{aligned} y^* &= y - XCr, \\ X^* &= XB. \end{aligned}$$

Искомые оценки находим из оценок МНК  $\hat{\beta}^*$  модифицированной регрессии

$$\hat{\beta}_r = B\hat{\beta}^* + Cr.$$

Бывает, что ограничения заданы не в явном виде, а как раз в виде линейного уравнения, связывающего некие «модифицированные» коэффициенты с исходными коэффициентами, т.е. как

$$\beta = B\beta^* + b.$$

Тогда выкладки упрощаются. Переменные модифицированной регрессии равны

$$\begin{aligned} y^* &= y - Xb, \\ X^* &= XB, \end{aligned}$$

а оценки МНК с ограничениями рассчитываются как

$$\hat{\beta}_r = B\hat{\beta}^* + b,$$

где  $\hat{\beta}^* = (X^{*\top} X^*)^{-1} X^{*\top} y$ .

Типичный пример — модель полиномиального лага (лага Олмон).

.....

### Проверка статистических гипотез

Пусть данные распределены как  $F_{\theta}$ , где вектор параметров  $\theta$  принадлежит некоторому множеству  $\Theta$ . Проверяется гипотеза, что на самом деле вектор параметров  $\theta$  принадлежит более узкому множеству  $\Theta_0$ ,  $\Theta_0 \subset \Theta$ . На практике обычно  $\theta \in \mathbb{R}^m$ , а множество  $\Theta_0$  задается уравнениями:

$$q(\theta) = 0_k.$$

Предположение о том, что  $\theta \in \Theta_0$  (т.е. о том, что выполнены равенства  $q(\theta) = 0_k$ ), называют нулевой гипотезой и обозначают  $H_0$ . Наряду с нулевой гипотезой рассматривают альтернативную гипотезу:  $\theta \in \Theta_A$ , где  $\Theta_A \subset \Theta$ ,  $\Theta_A \cap \Theta_0 = \emptyset$ . Альтернативную гипотезу обозначают  $H_A$ . Чаще всего  $\Theta_A = \Theta \setminus \Theta_0$ , т.е. при нулевой гипотезе  $q(\theta) = 0_k$  альтернативная гипотеза состоит в том, что  $q(\theta) \neq 0_k$ .

В простейшем случае нулевая гипотеза состоит в том, что выполнены линейные ограничения, а альтернативная — в том, что они не выполнены, т.е.

$$\begin{aligned} H_0: & \quad R\theta = r, \\ H_A: & \quad R\theta \neq r. \end{aligned}$$

Это так называемая линейная гипотеза.

Процедуру (правило) проверки гипотезы называют критерием.

Проверяют гипотезу обычно с помощью некоторой статистики  $s$ . Под статистикой понимают функцию данных:  $s = s(\mathbf{x})$ , где  $\mathbf{x}$  — данные. Обычно  $s$  — скалярная величина.

Необходимо, чтобы распределение статистики в предположении, что верна нулевая гипотеза, было известным. Распределение  $s$ , вообще говоря, может зависеть от неизвестных истинных параметров  $\theta$ . В качестве статистик, с помощью которых проверяются гипотезы, имеет смысл использовать только такие статистики, распределения которых не зависят от  $\theta$ , либо такие, для которых эта зависимость асимптотически исчезает.

Процедура проверки следующая: если  $s$  меньше некоторой критической границы  $s^*$ , то  $H_0$  принимается, а если, наоборот, то  $H_0$  отвергается (и, соответственно, принимается  $H_A$ ):

$$\begin{aligned} s > s^* & \text{ — принимаем } H_0, \\ s < s^* & \text{ — принимаем } H_A. \end{aligned}$$

(Мы подразумеваем, что  $s$  имеет непрерывное распределение. В таком случае событие  $s = s^*$  можно не рассматривать, так как его вероятность равна нулю.)

Условие  $s > s^*$  задает правый хвост распределения статистики  $s$ . Вероятность попадания в этот хвост выбирается малой. Вероятность рассчитывается в предположении, что верна  $H_0$ . Обозначим эту вероятность  $\alpha^*$ :

$$\alpha^* = \text{Prob}(s > s^* | H_0).$$

Чаще всего берут  $\alpha^* = 5\%$  (иногда  $\alpha^* = 1\%$  или  $\alpha^* = 10\%$ ), хотя каких-то особых оснований для этого нет.

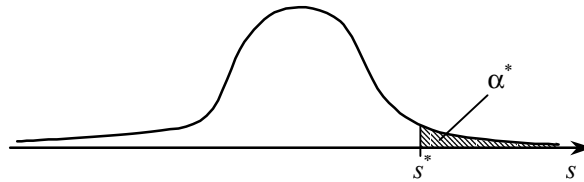


Рисунок 1

Если  $H_0$  не верна, а верна  $H_A$ , то распределение  $s$  будет каким-то другим.

Если параметр  $\theta \in \Theta_A$ , в соответствии с которым порождаются данные, достаточно далек от попадания в  $\Theta_0$ , и мы удачно выбрали статистику  $s$ , то ее распределение сместится

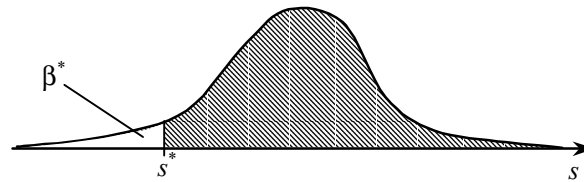


Рисунок 2

так, что вероятность попадания в хвост  $s > s^*$  станет велика.

Если верна  $H_0$ , то с вероятностью  $\alpha^*$  мы совершим ошибку, отвергнув верную гипотезу  $H_0$ . Это так называемая ошибка 1-го рода. Если же  $H_0$  неверна, а верна  $H_A$ , то с некоторой вероятностью  $\beta^*$  мы совершим противоположную ошибку — примем неверную гипотезу  $H_0$ . Это ошибка 2-го рода.

Вероятность ошибки 1-го рода ( $\alpha^*$ ) называют размером или уровнем значимости критерия. Один минус вероятность ошибки 2-го рода ( $1 - \beta^*$ ) называют мощностью критерия.

Размер  $\alpha^*$  выбирают обычно на каком-то фиксированном уровне (например,  $\alpha^* = 0,05$ ).

Если критерий не подходит для проверки рассматриваемой гипотезы, то распределение статистики при некотором параметре  $\theta$  из  $\Theta_A$  может сместиться влево, а не вправо. При

этом мощность критерия окажется меньше уровня значимости ( $1 - \beta^* < \alpha^*$ ); такой критерий называют смещенным.

Если мы сдвигаем критическую границу вправо, то уровень значимости уменьшается, а мощность увеличивается при любых данных значениях параметров, соответствующих нулевой и альтернативной гипотезам. Общее правило, таким образом, состоит в том, что уровень значимости и мощность находятся в обратном соотношении.

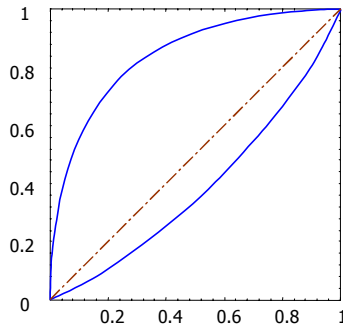


Рисунок 3

Следующий рисунок показывает типичный график зависимости мощности от уровня значимости. На этом графике кривая для несмещенного критерия должна лежать выше диагонали.

Вообще говоря, можно использовать любую статистику для проверки любой альтернативной гипотезы. Однако разные критерии имеют разную мощность против конкретного вида альтернативной гипотезы. Критерий бесполезен, если его мощность низка, а тем более, если он является смещенным. Желательно, чтобы мощность  $1 - \beta^*$  была как можно большей (мощный критерий = хороший критерий).

К сожалению, мощность, вообще говоря, зависит от неизвестных истинных параметров  $\theta$ :  $\beta^* = \beta^*(\theta)$ . Если  $\Theta_A$  состоит из единственной точки, то проблемы нет. В противном случае то, какой критерий лучше использовать, зависит от  $\theta$ . Если критерий является самым мощным при любом векторе  $\theta \in$

$\Theta_A$ , то такой критерий называют равномерно наиболее мощным критерием.

Важными являются также асимптотические свойства критерия. Желательно, чтобы при каждом фиксированном значении истинных параметров, не удовлетворяющих нулевой гипотезе, и при каждом фиксированном уровне значимости вероятность ошибки 2-го рода стремилась бы к нулю (а мощность, соответственно, стремилась бы к единице) по мере того, как количество наблюдений стремится к бесконечности. Это свойство называют состоятельностью критерия.

Следующий рисунок иллюстрирует понятие состоятельности. Крайняя левая кривая показывает распределение статистики в случае, когда выполнена нулевая гипотеза. Для упрощения иллюстрации взята такая статистика, распределение которой при нулевой гипотезе не зависит от количества наблюдений. Соответственно, критическая граница на рисунке не сдвигается при изменении количества наблюдений. Остальные три кривые показывают распределение той же статистики при некотором фиксированном векторе параметров, для которого не выполнена нулевая гипотеза. Кривые отличаются количеством наблюдений, по которым строится статистика. Распределение состоятельной статистики так же, как на рисунке, должно, как правило, смещаться вправо по мере увеличения количества наблюдений. При достаточно большом количестве наблюдений вероятность ошибки 2-го рода должна быть пренебрежимо мала.

Состоятельность критерия означает, что мы потенциально, имея бесконечно много наблюдений, можем выяснить с помощью данного критерия, удовлетворяет ли нулевой гипотезе тот вектор параметров, в соответствии с которым порождаются данные. Так же как и для состоятельности оценок,

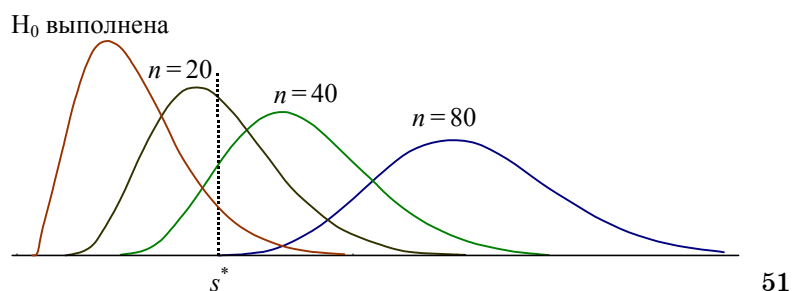


Рисунок 4

здесь следует сделать оговорку, что исследователь имеет каждый раз конкретную конечную выборку, и состоятельность критерия ему мало помогает.

Не всегда удобно пользоваться критической границей  $s^*$ . Можно ввести функцию  $\alpha(s)$ :

$$\alpha(\tilde{s}) = \text{Prob}(s > \tilde{s}).$$

$\alpha(s)$  — это уровень значимости, соответствующий границе  $s$ . Это статистика, так же как и  $s$ . Однако в отличие от обычно используемых статистик,  $H_0$  принимается, если  $\alpha(s)$  больше некоторой критической границы  $\alpha^*$ , и  $H_0$  отклоняется в противном случае:

$$\begin{aligned} \alpha(s) > \alpha^* & \text{ — принимаем } H_0, \\ \alpha(s) < \alpha^* & \text{ — принимаем } H_A. \end{aligned}$$

Удобство использования в качестве статистики состоит в том, что  $\alpha(s)$  при нулевой гипотезе имеет равномерное распределение и в качестве критической границы для нее берется непосредственно  $\alpha^*$  (например,  $\alpha^* = 0,05$ ). Хорошей практикой при изложении результатов проверки гипотезы является указание наряду с величиной исходной статистики  $s$  также ее уровня значимости  $\alpha(s)$ .

### Проверка линейных гипотез в регрессии

Мы хотим проверить гипотезу  $H_0: R\beta = r$ . Прежде, чем перейти к построению статистики, с помощью которой можно проверить данную гипотезу, докажем вспомогательное утверждение.

#### Теорема.

Пусть  $\hat{\beta}$ ,  $\hat{y}$ ,  $e$  — оценки коэффициентов, расчетные значения и остатки из регрессии  $y$  по  $X$ , а  $\hat{\beta}_R$ ,  $\hat{y}_R$ ,  $e_R$  — те же величины из регрессии  $y$  по  $X$  с ограничениями  $R\beta = r$ . Тогда имеет место следующее равенство:

$$\begin{aligned} \|\hat{y} - \hat{y}_R\|^2 &= \|e_R - e\|^2 = e_R^T e_R - e^T e = \\ &= (R\hat{\beta} - r)^T (R(X^T X)^{-1} R^T)^{-1} (R\hat{\beta} - r) = \\ &= (\hat{\beta} - \hat{\beta}_R)^T R^T (R(X^T X)^{-1} R^T)^{-1} R (\hat{\beta} - \hat{\beta}_R). \end{aligned}$$

#### Замечание.

Теорема опирается только на алгебраические свойства, и не использует никаких статистических предположений. Это просто тождество, которое является следствием того, что для оценивания двух рассматриваемых регрессий использовался МНК.

Доказательство:

Величину  $\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_R\|^2$  можно выразить также через величину невязки  $R\hat{\beta} - r$ . МНК с ограничениями, как показано выше, дает оценки

$$\hat{\beta}_R = \hat{\beta} - (X^T X)^{-1} R^T (R(X^T X)^{-1} R^T)^{-1} (R\hat{\beta} - r),$$

Домножим на  $X$  слева. Получим

$$\hat{\mathbf{y}} - \hat{\mathbf{y}}_R = X(X^T X)^{-1} R^T (R(X^T X)^{-1} R^T)^{-1} (R\hat{\beta} - r),$$

откуда

$$\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_R\|^2 = (R\hat{\beta} - r)^T (R(X^T X)^{-1} R^T)^{-1} (R\hat{\beta} - r).$$

Кроме того, поскольку  $r = R\hat{\beta}_R$ , то это же выражение можно записать через разность оценок МНК с ограничениями и без ограничений:

$$\begin{aligned} (R\hat{\beta} - r)^T (R(X^T X)^{-1} R^T)^{-1} (R\hat{\beta} - r) &= \\ &= (\hat{\beta} - \hat{\beta}_R)^T R^T (R(X^T X)^{-1} R^T)^{-1} R(\hat{\beta} - \hat{\beta}_R). \end{aligned}$$

Вектор  $\hat{\mathbf{y}} - \hat{\mathbf{y}}_R$  есть разница расчетных значений из регрессии без ограничений и из регрессии с ограничениями. Поскольку расчетные значения и остатки в сумме составляют зависимую переменную, то этот вектор равен также разности остатков  $e_R - e$ . Поэтому

$$\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_R\|^2 = \|e_R - e\|^2$$

Еще одно представление этой величины, которое может быть удобно в практических вычислениях:

$$\|e_R - e\|^2 = e_R^T e_R - e^T e = RSS_R - RSS.$$

Докажем это равенство. Поскольку  $\hat{\mathbf{y}} - \hat{\mathbf{y}}_R = e_R - e$ , то

$$e_R - e = X(X^T X)^{-1} R^T (R(X^T X)^{-1} R^T)^{-1} (R\hat{\beta} - r).$$

Пользуясь нормальным уравнением  $e^T X = 0$ , получим

$$e^T (e_R - e) = 0,$$

т.е.

$$e^T e = e^T e_R = e_R^T e.$$

Отсюда

$$\|e_R - e\|^2 = e_R^T e_R - e^T e_R - e_R^T e + e^T e = e_R^T e_R - e^T e. \quad \blacksquare$$

Для краткости обозначим величину, несколько альтернативных представлений которой приведено в данной теореме, через  $\delta$ . Воспользуемся величиной  $\delta$  для того, чтобы получить статистику для проверки гипотезы  $R\beta = r$ .

Выразим невязки через ошибки:  $\hat{\beta}$

$$\begin{aligned} R\hat{\beta} - r &= R(X^T X)^{-1} X^T y - r = R(X^T X)^{-1} X^T (X\beta + \epsilon) - r = \\ &= R\hat{\beta} - r + R(X^T X)^{-1} X^T \epsilon = R(X^T X)^{-1} X^T \epsilon. \end{aligned}$$

Заметьте, что здесь мы воспользовались предположением о том, что истинный вектор коэффициентов  $\beta$  удовлетворяет ограничениям:  $R\beta = r$ .

Отсюда

$$\begin{aligned} \hat{y} - \hat{y}_r &= X(X^T X)^{-1} R^T (R(X^T X)^{-1} R^T)^{-1} (R\hat{\beta} - r) = \\ &= X(X^T X)^{-1} R^T (R(X^T X)^{-1} R^T)^{-1} R(X^T X)^{-1} X^T \epsilon = S\epsilon, \end{aligned}$$

где

$$S = X(X^T X)^{-1} R^T (R(X^T X)^{-1} R^T)^{-1} R(X^T X)^{-1} X^T.$$

Поскольку, как несложно заметить,  $S$  — симметричная идемпотентная матрица ( $S^T = S$ ,  $S^2 = S$ ), то

$$\hat{y} - \hat{y}_r = S\epsilon \sim \mathcal{N}(\mathbb{0}_n, \hat{\sigma}^2 S),$$

и

$$\frac{\delta}{\hat{\sigma}^2} = \frac{1}{\hat{\sigma}^2} \|\hat{y} - \hat{y}_r\|^2 \sim \chi^2(\mathbb{0}_n, k),$$

В последнем соотношении мы воспользовались тем, что  $\text{rank}(S) = \text{rank}(R) = k$  ( $k$  — количество ограничений).

С точки зрения наглядности наиболее полезно представление через невязки:

$$\frac{\delta}{\hat{\sigma}^2} = \frac{1}{\hat{\sigma}^2} (R\hat{\beta} - r)^T (R(X^T X)^{-1} R^T)^{-1} (R\hat{\beta} - r) \sim \chi^2(\mathbb{0}_n, k).$$

Данная величина имеет указанное распределение только если  $R\hat{\beta} = r$ . Если же  $R\hat{\beta} \neq r$ , то распределение интересующей нас статистики смещается вправо (это будет так называемое нецентральное распределение хи-квадрат).

Чем больше невязки отклоняются от нуля, тем дальше вправо сдвигается распределение. Это обеспечивает высокую мощность данной статистики против альтернативной гипотезы  $H_A: R\hat{\beta} \neq r$ .

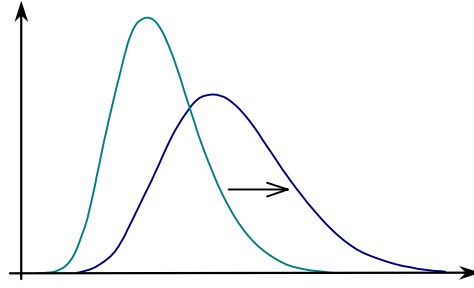


Рисунок 5

Однако величина дисперсии  $\sigma^2$  неизвестна. Выход из положения состоит в том, чтобы заменить дисперсию оценкой. Подходит несмещенная оценка из МНК без ограничений:

$$s^2 = \frac{\mathbf{e}^T \mathbf{e}}{n - m}.$$

Раньше мы получили, что

$$\frac{1}{\sigma^2} \mathbf{e}^T \mathbf{e} = \frac{RSS}{\sigma^2} \sim \chi^2(n - m).$$

Имеем две величины, распределенные как хи-квадрат —  $\frac{\delta}{\sigma^2}$  и  $\frac{RSS}{\sigma^2}$ . Если две независимые случайные величины имеют распределение хи-квадрат, то на их основе можно построить величину, которая имеет F-распределение Фишера. Поскольку  $\mathbf{e}_r - \mathbf{e} = \mathbf{S}\boldsymbol{\epsilon}$  и  $\mathbf{e} = \mathbf{M}\boldsymbol{\epsilon}$ , то

$$\mathbf{E}((\mathbf{e}_r - \mathbf{e})\mathbf{e}^T) = \mathbf{S}\mathbf{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)\mathbf{M} = \sigma^2 \mathbf{S}\mathbf{M} = \mathbf{0}.$$

Это означает, что  $\mathbf{e}$  и  $(\mathbf{e}_r - \mathbf{e})$  некоррелированы. Поскольку совместное распределение  $\mathbf{e}$  и  $(\mathbf{e}_r - \mathbf{e})$  является многомерным нормальным распределением, то, значит, они независимы. Значит, независимыми являются и функции от них,  $\delta = \|\mathbf{e}_r - \mathbf{e}\|^2$  и  $RSS = \|\mathbf{e}\|^2$ .

(Заметим, что  $\mathbf{e}_r$  и  $(\mathbf{e}_r - \mathbf{e})$  коррелированы. Поэтому мы не можем здесь взять вместо  $s^2$  несмещенную оценку дисперсии, основанную на регрессии с ограничениями, т.е.  $s_r^2 = \mathbf{e}_r^T \mathbf{e}_r / (n - m - k)$ .)

Эти рассуждения приводят к следующей статистике:

$$\frac{\delta/k}{e^T e / (n-m)} = \frac{\delta}{k s^2} \sim F_{k, n-m}$$

— статистика имеет F-распределение с  $k$  и  $(n-m)$  степенями свободы.

Наиболее удобная для практических целей запись этой статистики, по-видимому, следующая:

$$\frac{(RSS_r - RSS)/k}{RSS/(n-m)} \sim F_{k, n-m}.$$

Чтобы ее вычислить, достаточно знать суммы квадратов остатков в регрессии без ограничений и в регрессии с ограничениями.

Важно отметить, что важную роль при проверке ограничений на коэффициенты может играть матрица ковариаций оценок. Несмещенной оценкой ковариационной матрицы оценок является матрица

$$\hat{\text{Var}}(\hat{\beta}) = s^2 (X^T X)^{-1}.$$

Обозначим ее  $\hat{V}$ . Тогда полученную только что статистику можно переписать как

$$\begin{aligned} \frac{1}{k} \frac{\delta}{s^2} &= \frac{1}{k} \frac{(R\hat{\beta} - r)^T (R(X^T X)^{-1} R^T)^{-1} (R\hat{\beta} - r)}{s^2} = \\ &= \frac{1}{k} (R\hat{\beta} - r)^T (R\hat{V}R^T)^{-1} (R\hat{\beta} - r) \sim F_{k, n-m}. \end{aligned}$$

Удобство данной формулы состоит в том, что для вычисления статистики требуется знать информацию только для регрессии без ограничений. (Это так называемый принцип Вальда.)???

Для нелинейных гипотез вида  $q(\beta) = 0_k$ , где  $q(\cdot)$  — достаточно гладкая функция, можно использовать аналогичную формулу в качестве приближения:

$$\frac{1}{k} q(\hat{\beta})^T (R(\hat{\beta}) \hat{V} R(\hat{\beta})^T)^{-1} q(\hat{\beta}) \stackrel{a}{\sim} F_{k, n-m}.$$

Данную статистику называют статистикой Вальда. Здесь  $R(\hat{\beta})$  — матрица производных ограничений по параметрам:

$$R(\hat{\beta}) = \frac{\partial q(\hat{\beta})}{\partial \beta}.$$

В частном случае линейной гипотезы  $q(\beta) = R\beta - r$ , поэтому

$$R(\hat{\beta}) = \frac{\partial q(\hat{\beta})}{\partial \beta} = R.$$

Знак  $\overset{a}{\sim}$  следует читать «асимптотически имеет распределение». Асимптотические статистики широко используются в эконометрике при проверке гипотез, поскольку в нелинейных моделях и/или при проверке нелинейных ограничений, как правило, не удается получить точный закон распределения статистик.

Мы могли бы получить по аналогии с линейным случаем и другие асимптотические статистики для проверки нелинейных гипотез, например,

$$\frac{(RSS_R - RSS)/k}{RSS/(n-m)} \overset{a}{\sim} F_{k, n-m},$$

однако статистика Вальда является в данном случае (когда модель без ограничений линейна) наиболее удобной, поскольку для ее вычисления не требуется оценивать регрессию с нелинейными ограничениями.

Предположим, что мы одним из способов получили статистику  $F$ , имеющую F-распределение с  $k$  и  $(n-m)$  степенями свободы. Проверка гипотезы, как уже говорилось, обычно проводится одним из двух способов:

1) Вычисляем статистику  $F$ . Выбираем уровень значимости  $\alpha^*$  и находим по таблице F-распределения соответствующую критическую границу  $F^*$ . Если  $F < F^*$ , то принимаем нулевую гипотезу. Если  $F > F^*$ , то отклоняем нулевую гипотезу.

### Критерии удаления переменных

Пусть вектор параметров  $\beta$  состоит из двух частей —  $\beta_1$ ,  $\beta_2$ , и мы хотим проверить гипотезу о том, что  $\beta_2 = 0$ . Соответствующий F-критерий называют критерием удаления перемен-

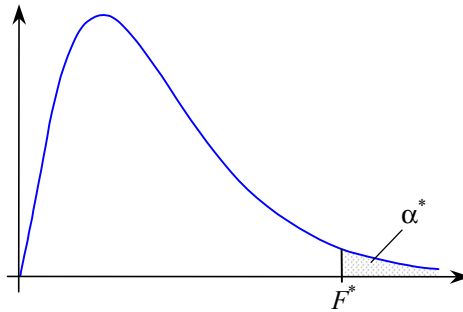


Рисунок 6

ных. Если в результате проверки гипотезы нулевая гипотеза отклоняется, то говорят, что рассматриваемые переменные статистически значимы — коэффициенты при них значимо отличаются от нуля.

Пусть первый регрессор является константой:  $x_1 = 1$ , и проверяется гипотеза о том, что  $\beta_2 = \dots = \beta_m = 0$ . В этом случае остатки в регрессии с ограничениями равны  $e_r = y^c$ , где  $y^c = y - \bar{y}$  — центрированная зависимая переменная. Статистика для проверки этой гипотезы имеет вид:

$$\frac{(y^c)^T y^c - RSS)/(m-1)}{RSS/(n-m)} \sim F_{m-1, n-m}.$$

2) Вычисляем статистику  $F$  и уровень значимости для этой статистики  $\alpha(F)$ . Выбираем пороговый уровень значимости  $\alpha^*$ . Если  $\alpha(F) < \alpha^*$ , то принимаем нулевую гипотезу. Если  $\alpha(F) > \alpha^*$ , то отклоняем нулевую гипотезу.

Проверяем гипотезу о том, что интересующие нас объясняющие переменные (т.е. все регрессоры, кроме константы) не имеют объяснительной силы. Получаем F-статистику для регрессии в целом. Если получена большая (превышающая критическую границу) статистика, то отвергаем нулевую гипотезу и считаем переменные значимыми.

Другой важный частный случай: значим ли отдельный регрессор. Проверяем гипотезу  $H_0: \beta_j = 0$ .

Воспользуемся статистикой

$$\frac{1}{k} (R\hat{\beta} - r)^T (R\hat{V}R^T)^{-1} (R\hat{\beta} - r) \sim F_{k, n-m}.$$

В данном случае имеем одно ограничение ( $k=1$ ) и

$$R = (0, \dots, 0, 1, 0, \dots, 0), \quad r = 0, \\ R\hat{\beta} - r = \hat{\beta}_j, \quad (R\hat{V}R^T)^{-1} = \hat{V}_{jj}^{-1}.$$

Здесь мы обозначили через  $\hat{V}_{jj}^{-1}$   $j$ -й диагональный элемент ковариационной матрицы параметров. Величина  $\hat{V}_{jj}^{-1}$  является оценкой дисперсии  $\hat{\beta}_j$ :

$$\hat{V}_{jj}^{-1} = \text{Var}(\hat{\beta}_j).$$

Она рассчитывается по формуле

$$\text{Var}(\hat{\beta}_j) = s^2 (\mathbf{X}^T \mathbf{X})_{jj}^{-1},$$

где  $(\mathbf{X}^T \mathbf{X})_{jj}^{-1}$  —  $j$ -й диагональный элемент матрицы  $(\mathbf{X}^T \mathbf{X})^{-1}$ .

Окончательно получаем следующую статистику:

$$\frac{\hat{\beta}_j^2}{\hat{\text{Var}}(\hat{\beta}_j)} \sim F_{k, n-m}.$$

— это квадрат оценки коэффициента делённый на оценку дисперсии этой оценки.

Как правило, используют корень из этой величины, который имеет t-распределение (распределение Стьюдента) с  $(n - m)$  степенями свободы:

$$t_j = \frac{\hat{\beta}_j}{\sqrt{\hat{\text{Var}}(\hat{\beta}_j)}} \sim t_{n-m}.$$

Эту статистику называют **t-статистикой**.  $\sqrt{\hat{\text{Var}}(\hat{\beta}_j)}$  есть **стандартная ошибка** для  $j$ -го коэффициента. Обозначим ее через  $se_j$ . Стандартная ошибка показывает, насколько точно мы оценили коэффициент  $\hat{\beta}_j$ . Стандартная ошибка измеряется в тех же единицах, что и соответствующий коэффициент. Таким образом,

$$t_j = \frac{\hat{\beta}_j}{se_j} \sim t_{n-m}.$$

Если оценка коэффициента мала по сравнению со стандартной ошибкой, то t-статистика мала, и делаем вывод, что коэффициент незначим (статистически незначимо отличается от нуля).

В рассматриваемом случае применяют **двусторонний критерий**.

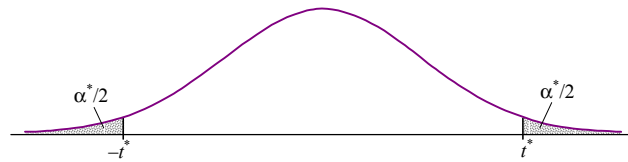


Рисунок 7

**рий.**

Если статистика  $t_j$  попадает в один из хвостов, то нулевую гипотезу отклоняют, т.е. коэффициент значим.

Фактически здесь в качестве статистики используется абсолютная величина t-статистики. Если  $|t_j| > t^*$ , то отвергаем  $H_0$ .

### Критерии правильности спецификации и критерий добавления переменной

Бывает удобно смотреть на проблему проверки гипотез с другой стороны — исходить из регрессии с ограничениями. Такой подход используется при проверке правильности спецификации модели, которую оценили. Идея состоит в следующем. Выбирается некая модифицированная модель, для которой исходная модель является частным случаем (т.е. модифицированная модель сводится к исходной, если ввести ограничения на параметры). Нулевая гипотеза состоит в том, что верна исходная модель. Если нулевая гипотеза отклоняется, значит, исходная модель неверно специфицирована, и ее следует усложнить. Описанный критерий можно назвать критерием правильности спецификации.

Рассмотрим только самый простой и часто встречающийся частный случай — критерий добавления переменной. Пусть  $X$  — матрица регрессоров в исходной модели. Мы хотим проверить, не был ли нами пропущен еще один регрессор —  $z$ . При этом рассматривается следующая модификация исходной модели:

$$y = X\beta + \gamma z + \epsilon.$$

По сути дела, соответствующая статистика была уже нами выведена выше, требуется только несколько изменить обозначения и найти оценку коэффициента при добавленной переменной ( $\hat{\gamma}$ ) через величины, получаемые при оценивании исходной модели. Последнюю проблему позволяет решить теорема о разбиении регрессоров. Согласно этой теореме,

$$\hat{\gamma} = (z^T M z)^{-1} y^T M z = \frac{z^T e}{z^T M z}.$$

Здесь, как и раньше,  $M = I - X(X^T X)^{-1} X^T$ . Мы воспользовались тем, что  $My = e$ . По той же теореме о разбиении регрессоров остатки в модифицированной регрессии  $e_A$  можно найти по формуле

$$e_A = My - \hat{\gamma} Mz = e - \frac{z^T e}{z^T M z} Mz.$$

Отсюда, пользуясь тем, что  $Me = e$ , имеем

$$e_A^T e_A = e^T e - 2\hat{\gamma} e^T z - \hat{\gamma}^2 z^T M z = e^T e - \frac{(z^T e)^2}{z^T M z}$$

или

$$RSS_A = RSS - \frac{(z^T e)^2}{z^T M z}.$$

Подставив это выражение в соответствующую F-статистику

$$\frac{(RSS - RSS_A)}{RSS_A / (n - m - 1)} \sim F_{1, n-m-1},$$

получим

$$\frac{1}{n - m - 1} \cdot \frac{(z^T e)^2}{RSS \cdot z^T M z - (z^T e)^2} \sim F_{1, n-m-1}.$$

Корень из этой величины распределен как t-Стьюдента с  $(n - m - 1)$  степенями свободы:

$$\frac{z^T e}{\sqrt{(n - m - 1)(RSS \cdot z^T M z - (z^T e)^2)}} \sim t_{n-m-1}.$$

Заметим, что  $z^T M z$  — сумма квадратов остатков в регрессии  $z$  по  $X$ . Существует простой алгоритм, позволяющий рассчитать эту величину. Пусть мы знаем ортонормированную матрицу  $\check{X}$ :  $\check{X} = X T^{-1}$  (см. описание ортогонализации Грама-Шмидта). Тогда

$$\begin{aligned} z^T M z &= z^T z - z^T X (X^T X)^{-1} X^T z = \\ &= z^T z - z^T \check{X} \check{X}^T z = \|z\|^2 - \|z^T \check{X}\|^2. \end{aligned}$$

Т.о., если мы только что рассчитали регрессию  $y$  по  $X$  с помощью метода ортогонализации, то мы сможем легко рассчитать эту статистику.

## Гипотезы, лежащие в основе МНК, и их невыполнение

Предположения МНК можно представить в виде пирамиды. Каждое последующее предположение рассматривают обычно только предполагая выполнение предыдущих. (Однако 2-е и 3-е предположение можно поменять местами).

①. Функциональная форма:

$$E(\epsilon) = E(y - X\beta) = 0.$$

②. Идентифицируемость:

②\*.  $X$  имеет полный ранг по столбцам.

②\*\*. Асимптотическая идентифицируемость.

Существует  $\text{plim}_{n \rightarrow \infty} [\frac{1}{n} \mathbf{X}^{(n)\top} \mathbf{X}^{(n)}] = \mathbf{M}$  и  $\mathbf{M}$  невырождена.

- ③. Ортогональность ошибок регрессорам
- ③\*. Матрица регрессоров  $\mathbf{X}$  детерминирована.
- ③\*\*. Матрица регрессоров  $\mathbf{X}$  и ошибки  $\boldsymbol{\varepsilon}$  независимы.
- ③\*\*\*.  $E(\boldsymbol{\varepsilon}|\mathbf{X}) = E(\boldsymbol{\varepsilon})$ .
- ③\*\*\*\*. Предел  $\text{plim}_{n \rightarrow \infty} [\frac{1}{n} \mathbf{X}^{(n)\top} \boldsymbol{\varepsilon}^{(n)}]$  существует и равен нулю:

$$\text{plim}_{n \rightarrow \infty} [\frac{1}{n} \mathbf{X}^{(n)\top} \boldsymbol{\varepsilon}^{(n)}] = \mathbf{0}_m.$$

- ④. Сферичность ошибок:

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \mathbf{I}.$$

$$\textcircled{4}^*. \boldsymbol{\varepsilon}_i \sim \text{IID}(0, \sigma^2).$$

- ⑤. Нормальность ошибок:

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}).$$

Часто предположения ① и ③ объединяют друг с другом. Так (об этом уже говорилось ранее при рассмотрении несмещенности) условия ① и ③\*\*\* в совокупности эквивалентны условию  $E(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}$ . Более того, только при выполнении условия ① условие ③ можно обозначить термином «ортогональность». Условие ③\*\*\*\* уже является в каком то смысле комбинацией предположений ① и ③.

Первые три предположения являются наиболее важными. Если они не выполнены, то использовать МНК для оценивания регрессионной модели не имеет смысла.

Кратко перечислим, какие свойства МНК следуют из каких предположений.

$$\textcircled{1} + \textcircled{2}^* + \textcircled{3}^{***} \Rightarrow \text{несмещенность}$$

$$\textcircled{2}^{**} + \textcircled{3}^{****} \Rightarrow \text{состоятельность}$$

$$\textcircled{1} + \textcircled{2}^* + \textcircled{3}^* + \textcircled{4} \Rightarrow \text{теорема Гаусса-Маркова}$$

$$\textcircled{1} + \textcircled{2}^* + \textcircled{3}^* + \textcircled{4} \Rightarrow \mathbf{s}^2(\mathbf{X}^\top \mathbf{X})^{-1} — \text{несмещенная оценка ковариационной матрицы оценок } \hat{\boldsymbol{\beta}}$$

$\textcircled{1} + \textcircled{2}^* + \textcircled{3}^* + \textcircled{4} + \text{некоторые дополнительные предположения} \Rightarrow \text{асимптотические распределения рассмотренных ранее величин такие же, как в предположении нормальности.}$

В том числе,

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} \mathcal{N}(\hat{\boldsymbol{\beta}}, \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}).$$

t-статистики для проверки линейных гипотез состоятельны и имеют асимптотически стандартное нормальное распределение.

F-статистики для проверки линейных гипотез состоятельны и имеют асимптотически распределение хи-квадрат.

① + ②\* + ③\* + ④ + ⑤  $\Rightarrow$  оценки МНК совпадают с оценками ММП, и, соответственно, являются асимптотически эффективными.

① + ②\* + ③\* + ④ + ⑤  $\Rightarrow$  верна стандартная теория проверки гипотез в конечных выборках (t- и F-статистики имеют точное распределение).

① + ②\* + ③\* + ④ + ⑤  $\Rightarrow$  оценки МНК имеют наименьшую дисперсию в классе несмещенных оценок.

### Функциональная форма

Предположим, что  $y$  порождается процессом вида

$$y_i = f_i(X_i, \theta) + \varepsilon_i \quad \text{при} \quad E(\varepsilon|X) = 0,$$

где  $f(X_i, \theta)$  не обязательно совпадает с той функциональной формой  $(X_i\beta)$ , которая оценивалась (функциональная форма другая, хотя факторы те же). Это называют неправильной спецификацией функциональной формы.

Тогда

$$\hat{\beta} = (X^T X)^{-1} X^T y = (X^T X)^{-1} X^T (\dot{y} + \varepsilon),$$

где  $\dot{y} = \{f(X_i, \theta)\}_i$ .

Понятно, что в общем случае коэффициенты  $\hat{\beta}$  не могут служить оценками параметров  $\theta$ . Кроме того, расчетные значения  $\hat{y} = X\hat{\beta}$  не могут служить оценками  $\dot{y}$ . Выразим расчетные значения через ошибки:

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T (\dot{y} + \varepsilon) = P(\dot{y} + \varepsilon),$$

где, как и раньше,  $P$  — матрица проекции на подпространство, натянутое на регрессоры.

Во-первых, эти оценки будут смещенными (рассмотрим только случай, когда  $\dot{y}$  и  $X$  детерминированы), т.е.

$$E(\hat{y}) = P\dot{y} \neq \dot{y},$$

за исключением случая, когда вектор  $\dot{y}$  лежит в подпространстве  $\mathcal{L}(X)$ , т.е. может быть представлен в виде  $\dot{y} = X\beta$  при каком-то векторе  $\beta$ .

Из этого следует также, что остатки из регрессии  $y$  по  $X$  в общем случае должны быть смещенными оценками исходных ошибок  $\epsilon$ .

Во-вторых, эти оценки будут несостоятельными. Действительно, если исходить из обычного предположения о том, что  $\text{plim}(P\epsilon) = 0$ , то  $\text{plim}(\hat{y}) = \hat{y}$  может быть выполнено только тогда, когда

$$\text{plim}(P\hat{y}) = \hat{y},$$

т.е. когда вектор  $\hat{y}$  лежит в подпространстве  $\mathcal{L}(X)$  асимптотически.

Аналогичным образом можно показать, что остатки из регрессии  $y$  по  $X$  в общем случае не могут быть состоятельными (в указанном выше смысле) оценками исходных ошибок  $\epsilon$ .

Типичный пример неправильной спецификации функциональной формы в рамках линейной модели — это случай пропущенных переменных.

Например, могут не быть учтены эффекты взаимодействия.

Пусть исходные факторы:

$$x_{i1}, \dots, x_{im}.$$

Эффектами второго порядка называют регрессоры следующего вида

$$x_{i1}^2, \dots, x_{i1}x_{im}, \dots, x_{ij}^2, \dots, x_{ij}x_{im}, \dots, x_{im}^2.$$

Имеем следующую модель:

$$y_i = \sum_{j=1}^m \beta_j x_{ij} + \sum_{j=1}^m \sum_{s=j}^m \gamma_{js} x_{ij} x_{is} + \epsilon_i.$$

Полный набор регрессоров здесь может быть линейно зависим — следует исключить лишние. Например, если  $x_{i1} = 1$ ,  $x_{i2} = z_i$ ,  $x_{i3} = z_i^2$ , то  $x_{i1}x_{i3}$  и  $x_{i2}^2$  — одно и то же.

Очевидно, что можно рассматривать и эффекты взаимодействия более высокого порядка.

Могут быть пропущены переменные, не являющиеся функциями исходных регрессоров.

Пусть данные порождены процессом вида

$$y = X_1\beta_1 + X_2\beta_2 + \epsilon,$$

однако при оценивании регрессоры  $X_2$  не были учтены.

Тогда

$$\hat{\beta}_1 = (X_1^T X_1)^{-1} X_1^T (X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \epsilon) =$$

$$= \mathring{\beta}_1 + (X_1^T X_1)^{-1} X_1^T X_2 \mathring{\beta}_2 + (X_1^T X_1)^{-1} X_1^T \epsilon,$$

Если регрессоры детерминированные, то

$$E(\tilde{\beta}_1) = \mathring{\beta}_1 + (X_1^T X_1)^{-1} X_1^T X_2 \mathring{\beta}_2.$$

Вектор  $\hat{\beta}_1$  является несмещенной оценкой  $\mathring{\beta}_1$  ( $E(\hat{\beta}_1) = \mathring{\beta}_1$ ) только когда матрицы  $X_1$  и  $X_2$  ортогональны:

$$X_1^T X_2 = O.$$

Как указывалось выше, остатки из регрессии  $y$  по  $X_1$  в общем случае должны быть смещенными оценками исходных ошибок  $\epsilon$ . Действительно,

$$e = y - X_1 \tilde{\beta}_1,$$

поэтому

$$\begin{aligned} E(e) &= E(y) - X_1 E(\tilde{\beta}_1) = \\ &= X_1 \mathring{\beta}_1 + X_2 \mathring{\beta}_2 - X_1 \mathring{\beta}_1 - X_1 (X_1^T X_1)^{-1} X_1^T X_2 \mathring{\beta}_2 = \\ &= (I - X_1 (X_1^T X_1)^{-1} X_1^T) X_2 \mathring{\beta}_2 = M_1 X_2 \mathring{\beta}_2, \end{aligned}$$

где  $M_1 = I - X_1 (X_1^T X_1)^{-1} X_1^T$ .

Особенно сильное смещение остатков наблюдается, когда  $X_1$  и  $X_2$  ортогональны, поскольку тогда  $M_1 X_2 = X_2$  и  $E(e) = X_2 \mathring{\beta}_2$ .

В частном случае, когда пропущена только одна переменная ( $z$ ), ортогональная  $X_1$ , имеем  $E(e) = \mathring{\beta}_2 z$ . Для обнаружения такой ошибки спецификации достаточно посмотреть на график  $e_i$  по  $z_i$ :

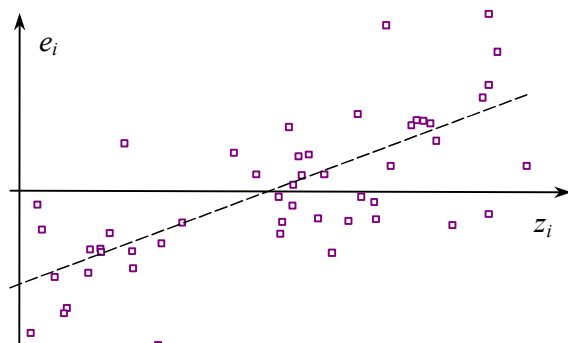


Рисунок 8

С другой стороны, если  $z$  почти линейно зависима от  $X_1$ , то на графике трудно заметить наличие зависимости.

Формальный метод обнаружения пропущенной переменной состоит в использовании соответствующего статистического критерия. Он был описан выше.

Претенденты на проверку:

- эффекты второго порядка (и более высоких порядков) и другие функции исходных переменных;
- расчетные значения и их различные степени (т.н. RESET Рамсея);
- если данные — временные ряды, то функции времени — тренды, фиктивные переменные;
- критерий Чоу — фиктивные переменные по подвыборкам;
- еще критерии для временных рядов — CUSUM и CUSUMSQ.

С другой стороны, можно посмотреть на пропущенные переменные как на случайные возмущения. Такой подход оправдан, если предположить, что

$$E(X_2 | X_1)\beta_2 = 0.$$

(Для этого достаточно, чтобы пропущенные переменные имели нулевое математическое ожидание  $E(X_2 | X_1) = 0$ .)

С этой точки зрения пропущенные переменные просто входят в ошибку (вообще говоря, ошибка — та же пропущенная переменная). Ошибкой тогда является величина  $X_2\beta_2$

+  $\varepsilon$ . Как несложно увидеть, при выполнении предположения  $E(X_2 | X_1)\beta_2 = 0$ , выполнено

$$E(X_2\beta_2 + \varepsilon | X_1)\beta_2 = 0,$$

поэтому оценки МНК  $\hat{\beta}_1 = (X_1^T X_1)^{-1} X_1^T y$  должны быть несмещенными оценками коэффициентов  $\beta_1$ . Также, при выполнении ряда стандартных предположений, эти оценки должны быть состоятельными.

В некоторых ситуациях естественно предположить, что строки матрицы пропущенных переменных  $X_2$  являются независимыми одинаково распределенными случайными векторами:

$$X_{i2} \sim \text{IID}$$

и при этом независимы от регрессоров  $X_1$ . В этом случае математическое ожидание их одно и то же. Если обозначить вектор математических ожиданий через  $\alpha$ , то

$$E(X_2 | X_1)\beta_2 = \alpha^T \beta_2 \cdot \mathbf{1}.$$

...Неэффективность оценки с точки зрения полной информации  $X_1, X_2$ . С другой стороны, полученные оценки эффективны (при стандартных предположениях) если доступна только информация об  $X_1$ .

Если в регрессии есть константа, то пропуск таких переменных приведет только к тому, что несостоятельной будет оценка константы.

Из этих рассуждений следует, что важно всегда включать в регрессию константу.

Однако, например, временной тренд нельзя считать случайной переменной — его пропуск может привести к смещению оценок и несостоятельности.

### Оценка с предварительной проверкой гипотезы

(pretest estimator)

Оценка с предварительной проверкой гипотезы

$$\beta_1^* = \begin{cases} \hat{\beta}_1^{(1)}, & F < F_\alpha \\ \hat{\beta}_1^{(1+2)}, & F > F_\alpha \end{cases}$$

Здесь  $F$  — F-статистика для гипотезы о том, что  $\beta_2 = 0$ ,  $F_\alpha$  — критическая граница.

Какая оценка более точная — не понятно. Это зависит от величины коэффициентов  $\hat{\beta}_2$  (от  $\text{Var}(X_2 \hat{\beta}_2)$ ) и от  $\sigma_\varepsilon^2$ .

Сравниваем  $\text{MSE}(\hat{\beta}_1^*)$ ,  $\text{MSE}(\hat{\beta}_1^{(1)})$  и  $\text{MSE}(\hat{\beta}_1^{(1+2)})$ . Аналитически трудно вычислить — используют метод Монте-Карло.

Не всегда следует спешить отбрасывать «незначимую» переменную.

.....

### Нарушение гипотезы об ортогональности

Может быть

$$y = X\beta + \varepsilon.$$

но при этом

$$E(\varepsilon | X) \neq 0.$$

Нарушение ортогональности.

Типичный случай — ошибки в переменных.

Пусть данные порождены процессом

$$y = X\beta + \varepsilon,$$

где

$$E(\varepsilon | X) \neq 0,$$

однако  $X$  не наблюдаются, вместо  $X$  наблюдается

$$X^* = X + \Psi.$$

(ошибки наблюдения  $\Psi$  независимы от  $\varepsilon$  и  $X$ )

$$E(\Psi | X) \neq 0 ???$$

$X^*$  тогда называют *proxу* для  $X$  — плохой заменитель, эрзац-переменная.

Тогда

$$y = X^*\beta + \varepsilon^*,$$

где

$$\varepsilon^* = \varepsilon - \Psi\beta = \varepsilon + (X - X^*)\beta.$$

$$E(\varepsilon^* | X^*) = -\Psi\beta = 0.$$

При дополнительных предположениях оказывается, что смещение сохраняется и асимптотически, т.е. оценки оказываются несостоятельными.

Вообще коррелированность между  $X$  и  $\varepsilon$  вызывает смещение (и, как правило, несостоятельность). ????

Еще один пример — «одновременность». См. далее тему «Системы одновременных уравнений».

На самом деле условие может быть ослаблено — оценки МНК останутся состоятельными:

$$E(\varepsilon_i | X_i, \dots, X_1) = 0.$$

Авторегрессия: регрессоры являются лагами (запаздывающими значениями) зависимой переменной.

— доказывается состоятельность.

Нарушение: авторегрессия и ошибки автокоррелированы — тогда несостоятельность.

$ARMA(1, 1)$ :

это авторегрессия 1-го порядка

$$y_i = \beta_1 + \beta_2 y_{i-1} + \varepsilon_i,$$

где ошибка представляет собой  $MA(1)$ -процесс:

$$\varepsilon_i = \xi_i + \mu \xi_{i-1}.$$

Тогда

$$E(\varepsilon_i | y_{i-1}, \dots, y_1) \neq 0,$$

так как ошибка  $\varepsilon_i$  зависит от  $\xi_{i-1}$  и  $y_{i-1}$  зависит от  $\xi_{i-1}$ .

Есть формальный критерий условия ортогональности (Дарбина-Ву-Хаусмана). Но об этом потом.???

## Идентифицируемость

$|X^T X| = 0$  — неполный ранг матрицы регрессоров, вырожденность. Однако чаще встречается почти вырожденность.

Данную проблему называют мультиколлинеарностью.

Оценки получаются неточными.

Пусть, например, для двух регрессоров,  $x_1$  и  $x_2$ , выполнено

$$x_1^T x_2 \ll \|x_1\| \cdot \|x_2\|$$

— трудно отличить влияние  $x_1$  на  $y$  от влияния  $x_2$  на  $y$ .

Причины: недостаточно данных, данные не варьируют.

Выход в экспериментальных науках: изменить регрессоры, например, сделав их ортогональными друг другу.

Показатели:

Число обусловленности ??

Как бороться?

Гребневая регрессия (ридж-регрессия) ???

Пусть у нас есть информация, что

$$E(\hat{\beta}) = 0, ???$$

тогда байесовская оценка:

$$\hat{\beta} = (X^T X + \delta I) X^T y.$$

### Сферичность

$$E(\epsilon \epsilon^T) = \sigma^2 I.$$

Предположение о сферичности ошибок обычно разбивают на две части: отсутствие автокорреляции ошибок и гомоскедастичность.

А. Автокорреляция ошибок (сериальная корреляция)

$$E(\epsilon_i \epsilon_{i-p}) \neq \sigma^2 I, \quad p = 1, 2, \dots$$

В. Мы предполагаем гомоскедастичность:

$$E(\epsilon_i^2) = \sigma^2, \quad \forall i = 1, \dots, n.$$

Нарушение этого предположения называют гетероскедастичностью:

$$E(\epsilon_i^2) = \sigma_i^2.$$

где не все  $\sigma_i^2$  одинаковы.

© оценки МНК  $\hat{\beta}$  являются смещенными при автокорреляции и наличии среди регрессоров лагов зависимой переменной;

©  $\sigma^2(X^T X)^{-1}$  — несостоятельная оценка ковариационной матрицы оценок  $\hat{\beta}$ ;

© обычные t- и F-статистики несостоятельны, их нельзя использовать для проверки гипотез;

© оценки МНК  $\hat{\beta}$  в предположении нормальности уже не будут эффективными, оценки менее точные, происходит потеря информации.

### Автокорреляция ошибок (сериальная корреляция)

Как выявить автокорреляцию?

Наиболее часто встречающийся вид автокорреляции ошибок — это автокорреляция первого порядка. Например, она возникает, когда ошибка порождена  $AR(1)$ -процессом:

$$\epsilon_i = \rho \epsilon_{i-1} + \xi_i.$$

В этом случае последовательные ошибки оказываются коррелированными.

Поскольку

$$E(\varepsilon_i \varepsilon_{i-1}) = \rho E(\varepsilon_{i-1}^2) + E(\varepsilon_{i-1} \xi_i) = \rho E(\varepsilon_{i-1}^2) = \rho \text{Var}(\varepsilon_i),$$

то

$$\text{Corr}(\varepsilon_i, \varepsilon_{i-1}) = \frac{E(\varepsilon_i \varepsilon_{i-1})}{\text{Var}(\varepsilon_i)} = \rho.$$

Таким образом, автокорреляция первого порядка не равна нулю. Аналогичным образом,

$$\text{Corr}(\varepsilon_i, \varepsilon_{i-p}) = \rho^p.$$

Наиболее известный критерий для выявления автокорреляции 1-го порядка — это [критерий Дарбина-Уотсона](#). Статистика для этого критерия имеет вид:

$$DW = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}.$$

Статистика всегда удовлетворяет неравенству  $0 < DW < 4$ .

Критерий предназначен для выявления автокорреляции только 1-го порядка, что можно увидеть из того, что данная статистика тесно связана с выборочной оценкой коэффициента автокорреляции 1-го порядка

$$\hat{\rho} = \frac{\sum_{i=2}^n e_i e_{i-1}}{\sum_{i=1}^n e_i^2}.$$

С точностью до пропуска первого и последнего наблюдения

$$DW \approx 2(1 - \hat{\rho}).$$

При отсутствии автокорреляции статистика  $DW$  должна быть приближенно равна двум. При положительной автокорреляции  $DW$  должна быть ближе к нулю. Точное распределение статистики зависит от матрицы регрессоров  $\mathbf{X}$ . Однако есть нижняя и верхняя границы для доверительной границы ( $d_\alpha^L$  и  $d_\alpha^U$ ), которые не зависят от  $\mathbf{X}$ . Именно они приводятся в таблицах распределения статистики Дарбина-Уотсона. (Есть компьютерные программы, которые считают точные критические границы.)

# Гипотезы, лежащие в основе МНК, и их невыполнение

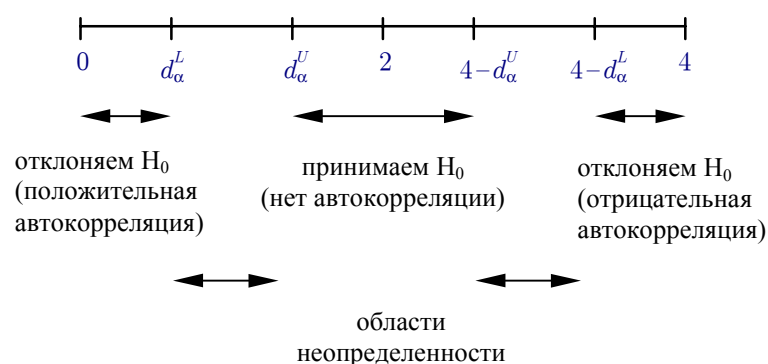


Рисунок 9. Интерпретация критерия Дарбина-Уотсона

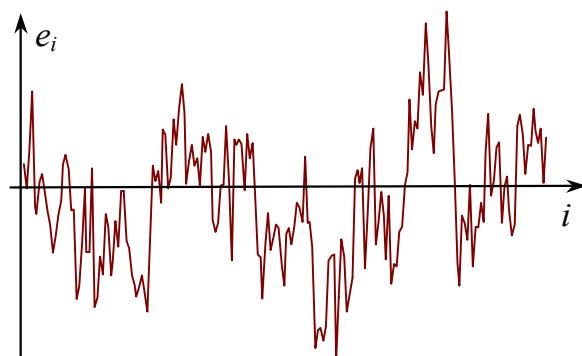


Рисунок 10. График остатков при сильной положительной коррелированности ошибок

Автокорреляцию ошибок можно также заметить на графике остатков:

При положительной автокорреляции за положительным остатком чаще следует положительный, а за отрицательным — отрицательный.

При отрицательной автокорреляции за положительным остатком чаще следует отрицательный остаток и наоборот.

Если среди регрессоров есть лаг зависимой переменной ( $y_{i-1}$ ), то критерий Дарбина-Уотсона оказывается несостоятельным. Статистику [h-Дарбина](#) применяют для выявления автокорреляции 1-го порядка в присутствии лага зависимой переменной.

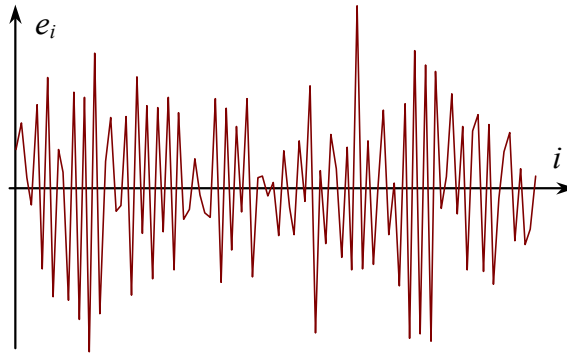


Рисунок 11. График остатков при сильной отрицательной коррелированности ошибок

Критерий Годфрея (альтернативный критерий Дарбина) применим при более общих предположениях.

Для расчета его статистики можно использовать вспомогательную регрессию:

$y$  от  $X$  и  $e_{-1}, \dots, e_{-p}$ .

— для автокорреляции  $p$ -го порядка.

Здесь  $e_{-j}$  — лаги остатков:

$$e_{-j} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ e_1 \\ \vdots \\ e_{n-j} \end{pmatrix} \Bigg\}^j .$$

Используем критерий добавления переменных.

Нулевая гипотеза во вспомогательной регрессии: добавочные переменные (лаги остатков) незначимы одновременно.

Процедура Кочрана-Оркатта

$AR(1)$ -процесс в ошибке

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

Инновации  $\xi_i$  не автокоррелированы, IID.

Сама регрессия имеет обычный вид

$$y_i = X_i \beta + \varepsilon_i.$$

Подставив в формулу  $AR(1)$ -процесса ошибку

$$\varepsilon_i = y_i - X_i \beta,$$

получим

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

Если бы мы знали  $\rho$ , то это была бы линейная регрессия и ее можно было бы оценить обычным образом, что можно увидеть из следующей записи:

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

Если использовать вместо истинного значения...

С другой стороны, если бы мы знали коэффициенты  $\beta$ , то это была бы линейная регрессия с неизвестным коэффициентом  $\rho$ .

Эти рассуждения приводят к итеративной процедуре:

0. Выбрать начальное приближение для  $\rho$ , например  $\rho = 0$ .

1. На основе оценки  $\rho$  ( $\hat{\rho}$ ) получить оценку вектора  $\beta$ .

Обозначим

$$\tilde{y}_{i-1} = y_i - \hat{\rho} y_{i-1} \quad \text{и} \quad \tilde{X}_{i-1} = X_i - \hat{\rho} X_{i-1} \quad (i = 1, \dots, n-1).$$

Тогда оценка  $\beta$  вычисляется по формуле

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

2. На основе оценок вектора  $\beta$  получить оценку коэффициента  $\rho$ :

$$\hat{\rho} = \frac{e^T e_{-1}}{e^T e},$$

где

$$e = y - X \hat{\beta}.$$

После этого снова переходим на шаг 1.

Это и есть метод Кочрана-Оркатта.

Заметим, что на 1-м шаге мы всегда получаем состоятельную оценку вектора  $\beta$ , если выполнены стандартное предположение о том, что ошибка в некотором статистическом смысле «ортогональна» матрице регрессоров.

Но если это не так, то полученная оценка может быть не состоятельной. Например, так будет, если среди регрессоров содержится лаг зависимой переменной ( $y_{t-1}$ ). В таком случае метод Кочрана-Оркатта неприменим.

На втором шаге мы получим состоятельную оценку  $\rho$ , поскольку она вычисляется на основе состоятельной оценки  $\hat{\beta}$ .

### Гетероскедастичность

Гетероскедастичность может быть разной. Предполагается, что дисперсия может быть некоторой функцией некоторой переменной  $z_i$ :

$$\sigma_i^2 = f(z_i).$$

Способ обнаружения.

Гетероскедастичность можно обнаружить графически —

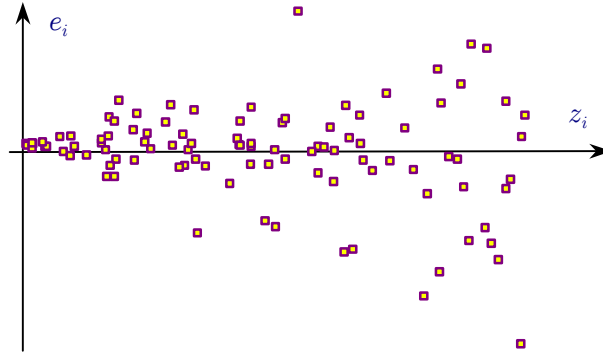


Рисунок 12

на графике остатков по «подозрительной» переменной  $z_i$ .

Так на приведенном графике видно, что при больших значениях  $z_i$  остатки более широко разбросаны, чем при малых.

#### Критерий Голдфелда-Кванда.

Пусть наблюдения разбиты на две группы, в первой —  $n_1$  наблюдений, во второй —  $n_2$ . (Считаем, что  $n_1 > m$  и  $n_2 > m$ .)

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \beta + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}.$$

Требуется проверить, что ошибка в обеих подвыборках имеет одну и ту же дисперсию.

Можно оценить регрессии по обеим выборкам по отдельности. В результате получим остатки  $e_1$  и  $e_2$ . Поскольку предполагается отсутствие корреляции между ошибками в подвыборках, то (в предположении нормальности ошибок) суммы квадратов, деленные на соответствующее количество степеней свободы, представляют собой две независимые ве-

личины, распределенные как хи-квадрат. Эти две величины представляют собой несмещенные оценки дисперсий в подвыборках —

$$s_1^2 = \frac{\mathbf{e}_1^T \mathbf{e}_1}{n_1 - m} \quad \text{и} \quad s_2^2 = \frac{\mathbf{e}_2^T \mathbf{e}_2}{n_2 - m}.$$

Деля одну из них на другую, получим F-статистику:

$$\frac{s_1^2}{s_2^2} \sim F_{n_1 - m, n_2 - m}.$$

Данную величину по понятной причине принято называть **дисперсионным отношением**. Смысл данной статистики состоит в том, что когда дисперсии в подвыборках сильно отличаются, то статистика будет либо существенно больше единицы, либо существенно меньше единицы. В данном случае естественно использовать двусторонний критерий. Это, конечно, не совсем обычно для критериев, основанных на F-статистике. Для уровня  $\alpha$  можно взять в качестве критических границ такие величины, чтобы вероятность попадания и в левый, и в правый хвост была одной и той же —  $\alpha/2$ . Заметим, что критические границы в этом случае будут обратными друг другу величинами, поскольку для F-распределения выполнено:

$$F(1 - \beta) = \frac{1}{F(\beta)}.$$

Нулевая гипотеза состоит в том, что нет гетероскедастичности (ошибки гомоскедастичны). Если дисперсионное отношение попадает в один из двух хвостов, то нулевая гипотеза отклоняется. В таком случае делается вывод, что ошибки гетероскедастичны.

Чтобы получить «двусторонний» уровень значимости, нужно умножить обычный «односторонний» уровень значимости на 2. Это верно, конечно, только для правого хвоста. Для левого хвоста нужно отнять «односторонний» уровень значимости от единицы и умножить на 2.

В основе данного критерия лежит предположение, что дисперсия ошибки зависит от переменной  $d$  следующего вида:

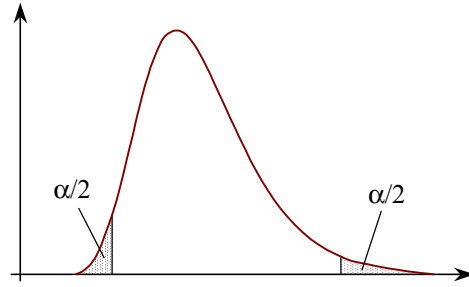


Рисунок 13. Двусторонний F-критерий

$$d_i = \begin{cases} 0, & i \leq n_1 \\ 1, & i > n_1 \end{cases}.$$

В общем случае дисперсия может зависеть от некоторой переменной  $z$ . В таком случае можно разбить наблюдения на две группы по некоторой границе  $z^*$ . Тогда формула для переменной  $d$  примет следующий вид:

$$d_i = \begin{cases} 0, & z_i \leq z^* \\ 1, & z_i > z^* \end{cases}.$$

Однако, рассматриваемый критерий должен иметь малую мощность против альтернативных гипотез такого более общего вида. Для увеличения мощности часто при расчете статистики не используют часть средних наблюдений (средних, если рассортировать наблюдения по порядку возрастания  $z_i$ ). При этом  $n_1 + n_2 < n$ .

Мощность критерия была бы больше, если бы в нем использовалась сама переменная  $z$ , от которой, как предполагается, может зависеть дисперсия. Но подобные критерии уже не будут иметь точного F-распределения в конечных выборках. Опишем один из наиболее простых критериев такого рода.

Статистику критерия можно получить с помощью вспомогательной регрессии:

$$e_i^2 \text{ от константы и } z_i.$$

Для проверки нулевой гипотезы используется обычная F-статистика из этой регрессии, соответствующая гипотезе о том, что коэффициент при  $z_i$  равен нулю. Конечно, эта статистика будет иметь F-распределение только приближенно, с асимптотической точки зрения.

В более общем случае таких «подозрительных» (с точки зрения влияния на дисперсию) переменных может быть несколько. Пусть они собраны в матрицу  $\mathbf{Z}$ . Вспомогательная регрессия в этом случае задается уравнением:

$$e_i^2 = \alpha + \mathbf{Z}_i \boldsymbol{\gamma} + \text{ошибка}.$$

Нулевая гипотеза о гомоскедастичности проверяется как гипотеза о том, что  $\boldsymbol{\gamma} = \mathbf{0}$ , т.е. что коэффициенты при всех переменных кроме константы равны нулю.

### Взвешенная регрессия

Пусть известно, что дисперсия пропорциональна некоторой переменной  $w_i$ :

$$\sigma_i^2 = \lambda w_i.$$

Поделив каждое наблюдение в исходной регрессии на  $\sqrt{w_i}$ , получим регрессию

$$\tilde{y}_i = \tilde{\mathbf{X}}_i \boldsymbol{\beta} + \tilde{\varepsilon}_i,$$

где

$$\tilde{y}_i = \frac{y_i}{\sqrt{w_i}}, \quad \tilde{\mathbf{X}}_i = \frac{1}{\sqrt{w_i}} \mathbf{X}_i, \quad \tilde{\varepsilon}_i = \frac{\varepsilon_i}{\sqrt{w_i}},$$

При этом в получившейся регрессии ошибки будут уже гомоскедастичны:

$$E(\tilde{\varepsilon} \tilde{\varepsilon}^T) = \lambda \mathbf{I}.$$

Таким образом, чтобы избавиться от гетероскедастичности, надо разделить каждое наблюдение на число, пропорциональное корню из дисперсии ошибки этого наблюдения.

Такая регрессия называется **взвешенной регрессией**. Числа  $w_i$  называют **весеами**. Здесь есть некоторая неоднозначность в терминологии. Иногда весами называют  $\sqrt{w_i}$  или  $1/w_i$ . Оценки коэффициентов во взвешенной регрессии являются решениями задачи минимизации взвешенной суммы квадратов:

$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^n \frac{1}{w_i} \varepsilon_i^2(\boldsymbol{\beta}) = \sum_{i=1}^n \frac{(y_i - \mathbf{X}_i \boldsymbol{\beta})^2}{w_i} \rightarrow \min \boldsymbol{\beta}.$$

Взвешенную регрессию можно рассчитать с помощью обычного эконометрического пакета как обычную регрессию, однако при этом  $R^2$  и F-статистика будут рассчитаны неверно, поскольку во взвешенной регрессии не будет константы. На месте константы ( $x_{i1} = 1$ ) в такой регрессии будет стоять переменная с типичным элементом

$$\tilde{x}_{i1} = \frac{1}{\sqrt{w_i}}.$$

Характерный пример использования взвешенной регрессии — регрессия с усредненными наблюдениями.

Предположим, что все исходные наблюдения  $\{1, \dots, n\}$  были разбиты на группы и нам известны не исходные данные, а только средние по группам. Обозначим множество наблюдений  $k$ -й группы через  $I_k$ , количество наблюдений в  $k$ -й группе через  $n_k$ . Мы предполагаем, что известны только величины

$$\bar{y}_k = \frac{1}{n_k} \sum_{i \in I_k} y_i, \quad \bar{X}_k = \frac{1}{n_k} \sum_{i \in I_k} X_i.$$

По сути дела, мы имеем дело с регрессией

$$\bar{y}_k = \bar{X}_k \beta + \bar{\varepsilon}_k,$$

где по аналогии

$$\bar{\varepsilon}_k = \frac{1}{n_k} \sum_{i \in I_k} \varepsilon_i \quad (\text{усредненные ошибки}).$$

Проблема состоит в том, что если исходные ошибки были гомоскедастичны, то усредненные ошибки в общем случае должны быть гетероскедастичны. Если исходные ошибки имели дисперсию  $\sigma^2$ , то усредненная ошибка для  $k$ -й группы имеет дисперсию

$$E(\bar{\varepsilon}_k^2) = \frac{1}{n_k^2} E\left[\left(\sum_{i \in I_k} \varepsilon_i\right)^2\right] = \frac{1}{n_k^2} \sum_{i \in I_k} E(\varepsilon_i^2) = \frac{1}{n_k^2} n_k \sigma^2 = \frac{\sigma^2}{n_k}.$$

Таким образом, дисперсия ошибки  $k$ -й группы обратно пропорциональна количеству наблюдений в группе. Поэтому для оценивания усредненной регрессии следует использовать взвешенную регрессию с весами  $w_k = 1/n_k$ , то есть домножить каждое усредненное наблюдение на корень из количества наблюдений в группе:

$$\sqrt{n_k} \bar{y}_k = \sqrt{n_k} \bar{X}_k \beta + \sqrt{n_k} \bar{\varepsilon}_k.$$

Заметим, что в регрессии с усредненными наблюдениями мы в общем случае не получим те же оценки, что и в исход-

ной регрессии. Более того, при усреднении теряется информация — оценки становятся менее точными в том смысле, что ковариационная матрица оценок усредненной регрессии не меньше, чем ковариационная матрица оценок исходной регрессии (т.е. их разность — положительно полуопределенная матрица). Совпадение оценок и ковариационных матриц возможно только когда в пределах каждой группы  $X_i$  одни и те же.

### Обобщенный МНК (метод Эйткена???)

Обобщенный МНК основан на той же идее, что и взвешенный МНК. По сути дела взвешенный МНК — это частный случай обобщенного МНК.

Пусть ковариационная матрица ошибок имеет следующий вид:

$$E(\mathbf{e}\mathbf{e}^T) = \lambda \mathbf{W},$$

где  $\mathbf{W}$  — известная матрица. Она должна быть симметричной и положительно определенной.

Коэффициенты  $\boldsymbol{\beta}$  в этом случае можно оценить по формуле

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

Поскольку  $\mathbf{W}$  — симметричная и положительно определенная матрица, то ее можно представить в виде

$$\mathbf{W} = \mathbf{A}^T \mathbf{A},$$

где  $\mathbf{A}$  — квадратная невырожденная матрица. Это может быть, например, разложение Холецкого. В таком случае матрица  $\mathbf{A}$  будет треугольной.

Обратная матрица  $\mathbf{W}^{-1}$  к  $\mathbf{W}$  будет иметь аналогичное представление:

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

Обобщенный МНК эквивалентен следующей вспомогательной регрессии:

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}} \boldsymbol{\beta} + \tilde{\mathbf{e}},$$

где

$$\tilde{\mathbf{y}} = (\mathbf{A}^{-1})^T \mathbf{y}, \quad \tilde{\mathbf{X}} = (\mathbf{A}^{-1})^T \mathbf{X}, \quad \tilde{\mathbf{e}} = (\mathbf{A}^{-1})^T \mathbf{e},$$

В этой регрессии ошибки уже будут сферическими:

$$E(\tilde{\mathbf{e}}\tilde{\mathbf{e}}^T) = \lambda \mathbf{I}.$$

(Следует предупредить, что, как и в частном случае взвешенной регрессии, если оценить вспомогательную регрессию с помощью обычного эконометрического пакета как обычную регрессию, то  $R^2$  и F-статистика будут рассчитаны неверно.)

Геометрически задача состоит в том, чтобы найти такой вектор  $\hat{y}$  из  $\mathcal{L}(X)$ , чтобы расстояние между  $y$  и  $\hat{y}$  было минимальным:

$$\begin{aligned} \|y - \hat{y}\| &\rightarrow \min \\ \hat{y} &\in \mathcal{L}(X). \end{aligned}$$

Однако это уже не евклидово расстояние:???

$$\|x' - x''\| = \sqrt{(x' - x'')^T W^{-1} (x' - x'')}$$

## Нормальность

Предположение о нормальности ошибок имеет вид:

$$\epsilon \sim \mathcal{N}(0, \sigma^2 I).$$

(Здесь, как и ранее, предположение о нормальности включает в себя также все предыдущие предположения.)

Если нормальность отсутствует, то не будут выполнены те свойства, о которых говорилось выше. Статистики  $t$  и  $F$  не будут иметь в конечных выборках точного  $t$ - и  $F$ -распределения, только асимптотически. Кроме того, оценки НК уже не будут оценками МП, и, следовательно, не будут асимптотически эффективными. Тем самым, если известно распределение ошибок, при использовании МНК теряется информация и оценки менее точны. Например, если распределение ошибок имеет толстые хвосты, то велика вероятность, что ошибка для какого-то наблюдения будет большой по абсолютной величине, и такое наблюдение сильно повлияет на оценки.

Если распределение ошибок точно известно, то вместо МНК можно использовать ММП с целью получения асимптотически эффективных оценок.

Если известно только, что распределение может иметь толстые хвосты, то вместо МНК можно использовать более робастные методы оценивания, то есть такие, которые менее чувствительны по отношению к отклонениям от нормальности. В частности, часто предлагается использовать метод наименьших отклонений (наименьших модулей). В нем минимизируется

не сумма квадратов остатков, а сумма модулей остатков. Целевая функция имеет вид:

$$\sum_{i=1}^n |\varepsilon_i(\beta)| = \sum_{i=1}^n |y_i - X_i \beta| \rightarrow \min_{\beta}.$$

Для того, чтобы оценить, насколько распределение отличается от нормального, обычно используют третий и четвертый центральный моменты

$$\mu_3 = E(\xi - E(\xi))^3, \quad \mu_4 = E(\xi - E(\xi))^4,$$

где  $\xi$  — случайная величина, имеющая данное распределение.

Моменты имеет смысл нормировать, чтобы можно было сравнивать распределения с разными дисперсиями. Тем самым, получаем коэффициенты асимметрии и куртозиса.

Асимметрия:

$$\frac{\mu_3}{\sigma^3}.$$

Куртозис:

$$\frac{\mu_4}{\sigma^4}.$$

Поскольку у нормального распределения куртозис равен 3, то применяют показатель, который равен нулю в случае нормального распределения. Этот коэффициент называют эксцессом.

Эксцесс:

$$\frac{\mu_4}{\sigma^4} - 3.$$

Распределения с положительным эксцессом обычно характеризуются более острой вершиной и более толстыми хвостами, чем нормальное распределение. Распределения с отрицательным эксцессом, наоборот, обычно характеризуются более плоской вершиной и более тонкими хвостами, чем нормальное распределение.

Наличие отклонений от нормальности можно заметить графически. Для этого можно использовать гистограмму, которая является оценкой плотности распределения. В данном случае используется гистограмма остатков, поскольку остатки являются оценками ошибок. Для того, чтобы на глаз заметить отклонения от нормальности, удобно наложить на гистограмму график плотности нормального распределения, с

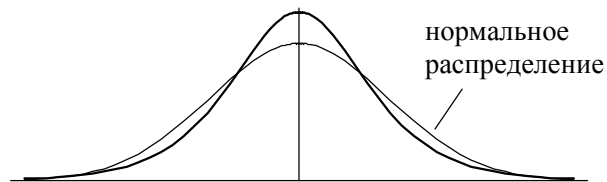


Рисунок 14. Распределение с положительным эксцессом по сравнению с нормальным распределением

такой же дисперсией, как у остатков (имеется в виду выборочная оценка дисперсии).

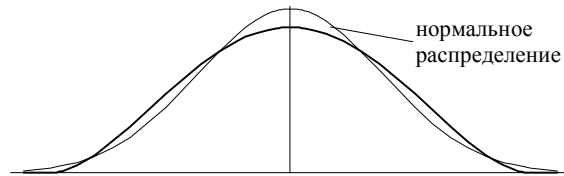


Рисунок 16. Распределение с отрицательным эксцессом по сравнению с нормальным распределением

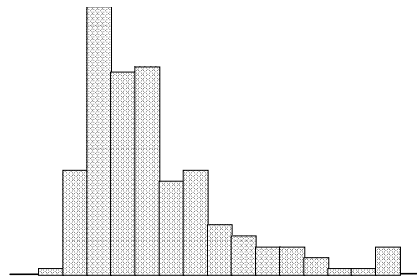


Рисунок 15. Гистограмма асимметричного распределения

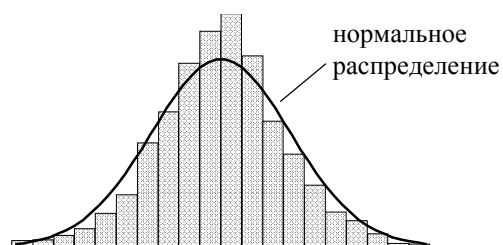


Рисунок 17. Гистограмма распределения с положительным эксцессом

То, что распределение ошибок имеет большой положи-

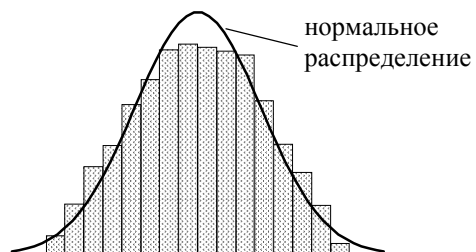


Рисунок 18. Гистограмма распределения с отрицательным эксцессом

тельный эксцесс, можно заметить также на графике остатков

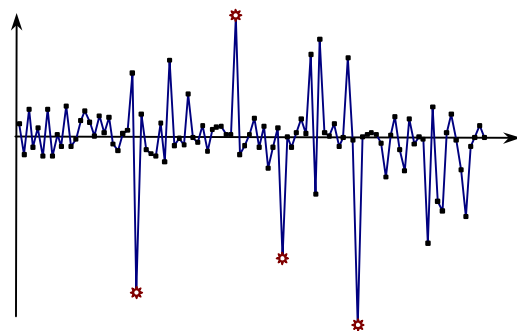


Рисунок 19. График остатков для распределения с большим положительным эксцессом. Выбросы отмечены звездочками

по какой либо переменной, например, по номеру наблюдения. На графике остатков в этом случае должно быть заметно большое количество выбросов, то есть сильно отклоняющихся наблюдений. Формальный критерий для проверки наличия выбросов мы рассмотрим в дальнейшем.

Для проверки нормальности можно использовать критерий Жарка-Беры. Статистика этого критерия основана на асимметрии и эксцессе.

Определим нормированные остатки как

$$\tilde{e}_i = \frac{e_i - \bar{e}}{s},$$

где  $\bar{e}$  — среднее остатков (оно равно 0 при наличии константы в регрессии),  $s$  — корень из  $s^2$ , оценки дисперсии.

Статистику критерия можно вычислить по формуле

$$\frac{1}{\sqrt{6n}} \sum_{i=1}^n \tilde{e}_i^3 + \frac{1}{\sqrt{24n}} \sum_{i=1}^n (\tilde{e}_i^4 - 3) \stackrel{a}{\sim} \chi^2(2).$$

Статистика имеет асимптотически распределение хи-квадрат с двумя степенями свободы. Первое слагаемое относится к асимметрии, второе — к эксцессу. Каждое из слагаемых асимптотически распределено как хи-квадрат с одной степенью свободы.

Следует сделать одно замечание. То, что при проверке нормальности кажется положительностью эксцесса распределения ошибок, на самом деле может быть следствием гетероскедастичности.

Пусть, например, дисперсия ошибки  $\sigma_i^2$  является случайной величиной. (В данном случае по смыслу это условная дисперсия.) Если условное распределение ошибок является нормальным,

$$\varepsilon_i | \sigma_i^2 \sim \mathcal{N}(0, \sigma_i^2),$$

то безусловное распределение должно в общем случае иметь положительный эксцесс. Гетероскедастичность такого рода (условная) не является гетероскедастичностью в обычном смысле слова (безусловной), поэтому она не приводит к несостоятельности оценок ковариационной матрицы оценок МНК. Однако как и с обычной гетероскедастичностью, если известна переменная, от которой зависит условная дисперсия, то использование МНК означает потерю информации.

Обычная, детерминированная, гетероскедастичность тоже может приводить к тому, что распределение ошибок будет казаться имеющим более острую вершину и более толстые хвосты, чем на самом деле. Из данных рассуждений можно сделать вывод, что следует сначала проверить предположение о гомоскедастичности, а затем уже проверять нормальность.