

Эконометрия I: регрессионный анализ

Курс эконометрии I состоит из двух частей: регрессионный анализ и временные ряды. Данное пособие предназначено для 1-й части курса, которая изучается в IV семестре.

Пособие включает 7 разделов:

1. Описательная статистика.
2. Случайные ошибки измерения.
3. Алгебра линейной регрессии.
4. Основная модель линейной регрессии.
5. Гетероскедастичность и автокорреляция ошибок.
6. Ошибки измерения факторов и фиктивные переменные.
7. Оценка параметров систем уравнений.

Каждый раздел открывается кратким обзором теоретического материала, затем следуют теоретические вопросы и задания, разбираемые на лекциях и семинарских занятиях, вслед за ними приводится набор задач и упражнений, которые решаются на практических занятиях и самостоятельно. Завершается каждый раздел списком литературы.

Теоретическая часть пособия подготовлена по материалам лекционного курса, прочитанного в 1992-96 гг., практическая часть в значительной мере построена по результатам работы по программе TESIS-TEMPUS в 1995-96 гг. .

Авторы: В.И. Суслов, Н.М. Ибрагимов, Б.Б. Карпенко, Е.А. Коломак.

1. Описательная статистика

1.1. Ряды наблюдений и их характеристики

$x_i, i = 1, \dots, N$ – ряд наблюдений за непрерывной случайной переменной x , вариационный ряд, выборка.

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i - \text{среднеарифметическое значение};$$

$$\hat{x}_i = x_i - \bar{x} - \text{центрированные значения наблюдений};$$

$$\frac{1}{N} \sum_{i=1}^N |\hat{x}_i| - \text{среднее линейное отклонение};$$

$x_{0.5}$ – медиана, т.е. среднее значение в ряду наблюдений:

если x_i упорядочены по возрастанию, то она равна $x_{\frac{N+1}{2}}$ при N нечетном и

$\left(\frac{x_N}{2} + \frac{x_{N+1}}{2}\right)$ при N четном;

$$m(q, c) = \frac{1}{N} \sum_{i=1}^N (x_i - c)^q - \text{моменты } q\text{-го порядка, центральные при}$$

$c = \bar{x}$, начальные при $c = 0$.

$$m(1, 0) = \bar{x},$$

$$m(2, \bar{x}) = \text{var}(x) = s^2, \text{ дисперсия } x,$$

s – среднеквадратическое (стандартное) отклонение,

$$\frac{\hat{x}_i}{s} - \text{центрированные и нормированные значения наблюдений},$$

$$\frac{s}{\bar{x}} - \text{коэффициент вариации},$$

$$m(3, \bar{x}) = m_3, m(4, \bar{x}) = m_4,$$

$$r_3 = \frac{m_3}{s^3} - \text{показатель асимметрии, если } r_3 \approx 0, \text{ то распределение величины}$$

симметрично, если $r_3 > 0$, то имеет место правая асимметрия, если $r_3 < 0$, - левая асимметрия;

$$r_4 = \frac{m_4}{s^4} - \text{показатель эксцесса (куртозиса), если } r_4 \approx 3, \text{ то распределение}$$

близко к нормальному, если $r_4 > 3$, то распределение высоковершинное, если $r_4 < 3$, - низкововершинное.

Пусть наряду с величиной x имеется N наблюдений y_i за величиной y .

$$m_{xy} = \text{cov}(x, y) = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) - \text{ковариация } x \text{ и } y,$$

$$r_{xy} = \frac{m_{xy}}{s_x s_y} - \text{коэффициент корреляции } x \text{ и } y; \quad -1 \leq r_{xy} \leq +1; \text{ если}$$

$r_{xy} \approx 0$, то величины x и y линейно независимы, если $r_{xy} \approx 1$, то они положительно линейно зависимы, если $r_{xy} \approx -1$, - отрицательно линейно зависимы.

1.2. Эмпирические распределения случайной величины

Пусть все $x_i, i = 1, \dots, N$ попадают в полуинтервал $(x_{0.0}, x_{1.0}]$, который делится на k равных полуинтервалов длиной Δ ; $x_{1.0} - x_{0.0} = k\Delta$. (предполагается, что $x_{0.0}$ “чуть” меньше или равно $\min x_i$, а $x_{1.0}$ “чуть” больше или равно $\max x_i$, так что некоторые из x_i попадают как в 1-й, так и в последний из этих k полуинтервалов).

$x_{1.0} - x_{0.0}$ – общий размах вариации.

$k = 1 + 3.322 \ln N$ – оптимальное соотношение между k и N (формула Стерджесса).

$(x_{l-1}, x_l]$ – l -й полуинтервал $l = 1, \dots, k$, где

$$x_0 = x_{0.0}, x_l = x_{l-1} + \Delta, l = 1, \dots, k, x_k = x_{1.0}.$$

w_l – доля общего количества наблюдений N , попавших в l -й полуинтервал – частоты, эмпирические оценки вероятностей попадания в данный полуинтервал;

$$\sum_{l=1}^k w_l = 1;$$

$$\bar{x}_l = x_{l-1} + \frac{\Delta}{2}, l = 1, \dots, k - \text{центры полуинтервалов};$$

$F_l, l = 0, 1, \dots, k$ – накопленные частоты (эмпирические вероятности, с которыми значения величины в выборке не превышают x_l):

$$F_0 = 0, F_l = F_{l-1} + w_l, l = 1, \dots, k, F_k = 1;$$

$$f_l = \frac{w_l}{\Delta}, l = 1, \dots, k - \text{эмпирические плотности распределения вероятности.}$$

$$\bar{x} = \frac{1}{k} \sum_{l=1}^k \bar{x}_l w_l - \text{среднеарифметическое значение};$$

$$x_{0.5} = x_{l-1} + \frac{\Delta}{w_l} (0.5 - F_{l-1}) - \text{медиана, здесь } l\text{-й полуинтервал является}$$

медианным, т.е. $F_{l-1} < 0.5 < F_l$;

$$m(q, c) = \frac{1}{k} \sum_{l=1}^k (\bar{x}_l - c)^q w_l - \text{моменты } q\text{-го порядка};$$

$$x_a = x_{l-1} + \frac{\Delta}{w_l} (a - F_{l-1}) - a\text{-й } (a100\text{-процентный}) \text{ квантиль, т.е.}$$

значение величины, которое не превышает в выборке с вероятностью a ; здесь l -й полуинтервал является квантильным, т.е. $F_{l-1} < a < F_l$ (x_l являются квантилями с $a = F_l$);

$$\bar{x}_a = \frac{1}{a} \left\{ \sum_{j=1}^{l-1} \bar{x}_j w_j + [x_{l-1} + \frac{\Delta}{2w_l} (a - F_{l-1})] (a - F_{l-1}) \right\} - \text{среднее по той}$$

(нижней) части выборки, которая выделяется a -м квантилем (l -й полуинтервал также квантильный).

Среди квантилей особое значение имеют те, которые делят выборку на равные части (иногда именно эти величины называют квартилями):

$x_{0.5}$ — медиана;

$x_{0.25}, x_{0.5}, x_{0.75}$ — квартили;

$x_{0.1}, x_{0.2}, \dots, x_{0.9}$ — децили;

$x_{0.01}, x_{0.02}, \dots, x_{0.99}$ — процентиля.

$x_{0.9} - x_{0.1}$ — децильный размах вариации (может быть также квартильным или процентильным);

$$\frac{\bar{x} - 0.9\bar{x}_{0.9}}{0.1\bar{x}_{0.1}} - \text{децильный коэффициент вариации (может быть медианным,}$$

квартильным или процентильным).

$$x^o = x_{l-1} + \Delta \frac{f_l - f_{l-1}}{(f_l - f_{l-1}) + (f_l - f_{l+1})} - \text{мода, т.е. наиболее вероятное}$$

значение величины в выборке; здесь l -й полуинтервал является модальным, f_l на нем достигает максимума; если этот максимум единственный, то распределение величины называется унимодальным; если максимума два - бимодальным; в общем случае - при нескольких максимумах - полимодальным.

Гистограмма - эмпирическая (интервальная) функция плотности распределения; имеет ступенчатую форму: на l -м полуинтервале ($l=1, \dots, k$) принимает значение f_l ;

Полигон - функция, график которой образован отрезками, соединяющими точки $(x_0, 0), (\bar{x}_1, f_1), \dots, (\bar{x}_k, f_k), (x_k, 0)$.

Гистограмма и полигон могут строиться непосредственно по весам w_l , если (как в данном случае) все полуинтервалы $(x_{l-1}, x_l], l=1, \dots, k$ имеют одинаковую длину.

Кумюлята - эмпирическая (интервальная) функция распределения вероятности, график которой образован отрезками, соединяющими точки $(x_l, F_l), l=0, 1, \dots, k$.

Огиа - то же, что и кумюлята, или (в традициях советской статистики) функция, обратная кумюляте.

1.3. Теоретические функции распределения случайной величины

\mathbf{X} - случайная величина,

\mathbf{z} - детерминированная переменная.

$\mathbf{F}(\mathbf{z}) = \mathbf{P}(\mathbf{x} \leq \mathbf{z})$ – функция распределения вероятности \mathbf{X} ;

$\mathbf{f}(\mathbf{z}) = \frac{d\mathbf{F}}{d\mathbf{z}}$ – функция плотности распределения вероятности \mathbf{X} ;

$$\int_{-\infty}^{+\infty} \mathbf{f}(\mathbf{z}) d\mathbf{z} = 1, \mathbf{F}(\mathbf{z}) = \int_{-\infty}^{\mathbf{z}} \mathbf{f}(\xi) d\xi,$$

$\bar{\mathbf{x}} = \mathbf{E}(\mathbf{x}) = \int_{-\infty}^{+\infty} \mathbf{z} \mathbf{f}(\mathbf{z}) d\mathbf{z}$ – математическое ожидание, среднее (теоретическое);

$\mu(\mathbf{q}, \mathbf{c}) = \mathbf{E}[(\mathbf{x} - \mathbf{c})^{\mathbf{q}}] = \int_{-\infty}^{\infty} (\mathbf{z} - \mathbf{c})^{\mathbf{q}} \mathbf{f}(\mathbf{z}) d\mathbf{z}$ – моменты \mathbf{q} -го порядка (теоретические);

$\mu(2, \bar{\mathbf{x}}) = \sigma^2$ – дисперсия (теоретическая);

$\mu(3, \bar{\mathbf{x}}) = \mu_3, \mu(4, \bar{\mathbf{x}}) = \mu_4$;

$\rho_3 = \frac{\mu_3}{\sigma^3}$ – показатель асимметрии (теоретический),

$\rho_4 = \frac{\mu_4}{\sigma^4}$ – показатель эксцесса, куртозиса (теоретический).

Для квантиля \mathbf{x}_a : $\int_{-\infty}^{\mathbf{x}_a} \mathbf{f}(\mathbf{z}) d\mathbf{z} = \mathbf{a}$; для моды \mathbf{x} : максимум $\mathbf{f}(\mathbf{z})$ достигается

при $\mathbf{z} = \mathbf{x}$.

Если распределение случайной величины симметрично, то $\mathbf{f}(\mathbf{z}) = \mathbf{f}(-\mathbf{z})$ и $\mathbf{x}_a = -\mathbf{x}_{1-a}$. В этом случае можно использовать понятие двустороннего квантиля

\mathbf{x}_a , для которого $\int_{-\mathbf{x}_a}^{\mathbf{x}_a} \mathbf{f}(\mathbf{z}) d\mathbf{z} = \mathbf{a}$, и значение которого совпадает с $\mathbf{x}_{\frac{1+\mathbf{a}}{2}}$ - значением

обычного (одностороннего) квантиля.

Если распределение случайной величины унимодально, то в случае симметричности $\bar{\mathbf{x}} \approx \mathbf{x}_{0.5} \approx \mathbf{x}$, при правой асимметрии $\bar{\mathbf{x}} > \mathbf{x}_{0.5} > \mathbf{x}$, при левой асимметрии $\bar{\mathbf{x}} < \mathbf{x}_{0.5} < \mathbf{x}$.

1.4. Функции распределения, используемые в эконометрии

В силу центральной предельной теоремы математической статистики, ошибки измерения и “остатки”, необъясняемые “хорошей” эконометрической моделью, имеют распределения близкие к нормальному. Поэтому все распределения, используемые в классической эконометрии, основаны на нормальном.

Пусть ϵ - случайная величина, имеющая нормальное распределение с нулевым мат.ожиданием и единичной дисперсией ($\epsilon \sim N(0,1)$). Функция плотности распределения ее прямо пропорциональна $e^{-\frac{\epsilon^2}{2}}$ (для наглядности в записи функции плотности вместо \mathbf{z} использован символ-имя самой случайной величины); **95**-процентный двусторонний квантиль $\epsilon_{0,95}$ равен **1.96**, **99**-процентный квантиль - **2.57**.

Пусть теперь имеется \mathbf{k} таких взаимно независимых величин $\epsilon_1 \sim N(0,1)$. Сумма их квадратов $\sum_{i=1}^k \epsilon_i^2$ является случайной величиной, имеющей распределение χ^2 с \mathbf{k} степенями свободы (обозначается χ_k^2). **95**-процентный (односторонний) квантиль $\chi_{k,0.95}^2$ при $\mathbf{k}=1$ равен **3.84** (квадрат **1.96**), при $\mathbf{k}=5$ - **11.1**, при $\mathbf{k}=20$ - **31.4**, при $\mathbf{k}=100$ - **124.3**.

Если две случайные величины ϵ и χ_k^2 независимы друг от друга, то случайная величина $\frac{\epsilon}{\sqrt{\chi_k^2/k}}$ имеет распределение t -Стьюдента с \mathbf{k} степенями свободы (t_k). Ее функция распределения прямо пропорциональна $(1 + \frac{t_k^2}{k})^{-\frac{k+1}{2}}$; в пределе при $\mathbf{k} \rightarrow \infty$ она становится нормально распределенной. **95**-процентный двусторонний квантиль $t_{k,0.95}$ при $\mathbf{k}=1$ равен **12.7**, при $\mathbf{k}=5$ - **2.57**, при $\mathbf{k}=20$ - **2.09**, при $\mathbf{k}=100$ - **1.98**.

Если две случайные величины $\chi_{k_1}^2$ и $\chi_{k_2}^2$ не зависят друг от друга, то случайная величина $\frac{\chi_{k_1}^2/k_1}{\chi_{k_2}^2/k_2}$ имеет распределение F -Фишера с \mathbf{k}_1 и \mathbf{k}_2 степенями свободы (F_{k_1, k_2}). **95**-процентный (односторонний) квантиль $F_{1, k_2, 0.95}$ при $\mathbf{k}_2=1$ равен **161**, при $\mathbf{k}_2=5$ - **6.61**, при $\mathbf{k}_2=20$ - **4.35**, при $\mathbf{k}_2=100$ - **3.94** (квадраты соответствующих $t_{k,0.95}$); квантиль $F_{2, k_2, 0.95}$ при $\mathbf{k}_2=1$ равен **200**, при $\mathbf{k}_2=5$ - **5.79**, при $\mathbf{k}_2=20$ - **3.49**, при $\mathbf{k}_2=100$ - **3.09**; квантиль $F_{k_1, 20, 0.95}$ при $\mathbf{k}_1=3$ равен **3.10**, при $\mathbf{k}_1=4$ - **2.87**, при $\mathbf{k}_1=5$ - **2.71**, при $\mathbf{k}_1=6$ - **2.60**.

Теоретические вопросы и задания

1. $\bar{x}(k) = \left(\frac{1}{N} \sum_{i=1}^N x_i^k \right)^{\frac{1}{k}}$ - среднее степенное.

При $k = -1$ это - среднее гармоническое,

при $k = 1$ - среднее арифметическое,

при $k = 2$ - среднее квадратическое.

Доказать, что

- $\bar{x}(k)$ растет с ростом k , равно $\min(x_i)$ при $k \rightarrow -\infty$ и $\max(x_i)$ при $k \rightarrow +\infty$;

- при $k = 0$ это - среднее геометрическое.

2(*). Для случая эмпирического распределения вывести формулы расчета среднего квантильного (\bar{x}_a), децильного коэффициента вариации и моды.

2. Случайные ошибки измерения

2.1. Первичные измерения

Пусть имеется N измерений x_i , $i = 1, \dots, N$ случайной величины x . Это - наблюдения за случайной величиной. Предполагается, что измерения проведены в неизменных условиях (факторы, влияющие на x , не меняют своих значений), и систематические ошибки измерений исключены. Тогда различия в результатах отдельных наблюдений (измерений) связаны только с наличием случайных ошибок:

$$x_i = \beta + \varepsilon_i,$$

где β - истинное значение x ,

ε_i - случайная ошибка в i -м наблюдении.

Если \mathbf{X} и $\mathbf{\varepsilon}$ - вектора-столбцы, соответственно, x_i и ε_i , а $\mathbf{1}_N$ - N -компонентный вектор-столбец, состоящий из единиц, то данную модель можно записать в матричной форме:

$$\mathbf{X} = \mathbf{1}_N \beta + \mathbf{\varepsilon}.$$

Предполагается, что ошибки по наблюдениям не зависят друг от друга и $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$, а их дисперсии по наблюдениям одинаковы $\text{var}(\varepsilon_i) = \sigma^2$, $i = 1, \dots, N$, или в матричной форме $\mathbf{E}(\mathbf{\varepsilon}\mathbf{\varepsilon}') = \mathbf{I}_N \sigma^2$ (где \mathbf{I}_N - единичная матрица размерности N). Требуется найти \mathbf{b} и \mathbf{e} - оценки, соответственно, β и ε_i . Для этого используется метод наименьших квадратов (МНК), т.е. искомые оценки определяются так, чтобы $\sum_{i=1}^N (x_i - b)^2 = \sum_{i=1}^N e_i^2 \rightarrow \min$ или $\mathbf{e}'\mathbf{e} \rightarrow \min$, где \mathbf{e} вектор-столбец оценок \mathbf{e}_i . В результате,

$$\mathbf{b} = \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i = \frac{1}{N} \mathbf{1}_N' \mathbf{X}, \quad \mathbf{e} = \mathbf{X} - \mathbf{1}_N \mathbf{b},$$

т.е. МНК-оценкой истинного значения измеряемой величины является среднее арифметическое по наблюдениям. Оценка \mathbf{b} относится к классу линейных, поскольку линейно зависит от наблюдений за случайной величиной.

В рамках сделанных предположений доказывается, что

- \mathbf{b} является несмещенной оценкой β ($\mathbf{b} = \mathbf{E}(\beta)$), ее дисперсия σ_b^2 равна $\frac{1}{N} \sigma^2$ и является минимальной на множестве линейных оценок; класс таких оценок (процедур оценивания) называют **BLUE** - **B**est **L**inear **U**nbiased **E**stimators;

- несмещенной оценкой σ^2 является

$$\hat{s}^2 = \frac{N}{N-1} s^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - b)^2 = \frac{1}{N-1} \mathbf{e}'\mathbf{e}.$$

Пусть теперь ϵ_i распределены нормально, тогда оценка максимального правдоподобия β совпадает с b , она несмещена, состоятельна (в пределе при $N \rightarrow \infty$ совпадает с β и имеет нулевую дисперсию) и эффективна (имеет минимально возможную дисперсию), величина $\frac{(b - \beta)\sqrt{N}}{\sigma}$ имеет распределение

$N(0,1)$ и $(1-\theta)100$ -процентный доверительный интервал для β определяется как

$$b \pm \frac{\sigma}{\sqrt{N}} \epsilon_{1-\theta},$$

где $\epsilon_{1-\theta}$ - $(1-\theta)100$ -процентный двусторонний квантиль нормального распределения.

Эта формула для доверительного интервала используется, если известно точное значение σ .

На практике точное значение σ , как правило, неизвестно, и используется другой подход.

Величина $\frac{(b - \beta)\sqrt{N}}{\hat{s}}$ имеет распределение t_{N-1} и $(1-\theta)100$ -процентный

доверительный интервал для β строится как

$$b \pm \frac{\hat{s}}{\sqrt{N}} t_{N-1, 1-\theta},$$

где $t_{N-1, 1-\theta}$ - $(1-\theta)100$ -процентный двусторонний квантиль t_{N-1} -распределения.

Поскольку величина β детерминирована, доверительные интервалы интерпретируются следующим образом: если процедуру построения доверительного интервала повторять многократно, то $(1-\theta)100$ процентов полученных интервалов будут содержать истинное значение β измеряемой величины.

2.2. Производные измерения

Пусть x_j , $j = 1, \dots, n$ - выборочные (фактические) значения (наблюдения, измерения) n различных случайных величин, β_j - их истинные значения, ϵ_j - ошибки измерений. Если \mathbf{x} , β , ϵ - соответствующие n -компонентные вектора-строки, то

$$\mathbf{x} = \beta + \epsilon.$$

Предполагается, что $E(\epsilon) = \mathbf{0}$, и ковариационная матрица ошибок $E(\epsilon \epsilon^T)$ равна Ω .

Пусть величина y рассчитывается как $f(\mathbf{x})$. Требуется найти дисперсию σ_y^2 ошибки $\epsilon_y = y - f(\beta)$ измерения (расчета) этой величины.

Разложение функции \mathbf{f} в ряд Тэйлора в фактической точке \mathbf{X} по направлению $\boldsymbol{\beta} - \mathbf{X}$ ($= -\boldsymbol{\varepsilon}$), если в нем оставить только члены 1-го порядка, имеет вид:

$$\mathbf{f}(\boldsymbol{\beta}) = \mathbf{y} - \boldsymbol{\varepsilon} \mathbf{g} \text{ или } \boldsymbol{\varepsilon}_y = \boldsymbol{\varepsilon} \mathbf{g} \text{ (заменяя “}\approx\text{” на “}=\text{”),}$$

где \mathbf{g} - градиент \mathbf{f} в точке \mathbf{X} (вектор-столбец с компонентами $g_j = \frac{\partial f}{\partial x_j}(\mathbf{x})$).

Откуда $\mathbf{E}(\boldsymbol{\varepsilon}_y) = \mathbf{0}$ и

$$\sigma_y^2 = \mathbf{E}(\mathbf{g}' \boldsymbol{\varepsilon} \boldsymbol{\varepsilon} \mathbf{g}) = \mathbf{g}' \boldsymbol{\Omega} \mathbf{g},$$

Это - общая формула, частным случаем которой являются известные формулы для дисперсии среднего, суммы, разности, произведения, частного от деления и др.

В случае, если ошибки величин x_j не скоррелированы друг с другом и имеют одинаковую дисперсию σ^2 ,

$$\sigma_y^2 = \mathbf{g}' \mathbf{g} \sigma^2.$$

В случае, если известны только дисперсии ошибок $\boldsymbol{\varepsilon}_j$, можно воспользоваться формулой, дающей верхнюю оценку дисперсии ошибки результата вычислений:

$$\sigma_y \leq \sum_{j=1}^n |\sigma_j g_j| = \Delta_y,$$

где σ_j - среднеквадратическое отклонение $\boldsymbol{\varepsilon}_j$.

Теоретические вопросы и задания

1. Проверить, что МНК-оценкой истинного значения измеряемой величины является среднее арифметическое по наблюдениям.

2(**). Доказать принадлежность этой оценки к классу **BLUE**.

3(**). Вывести формулу для дисперсии ошибки среднего.

4(**). Показать несмещенность оценки \hat{s}^2 дисперсии ошибки измерения.

5(*). Доказать, что в случае нормальности распределения ошибок измерения МНК-оценка истинного значения измеряемой величины совпадает с оценкой максимального правдоподобия.

6(**). Доказать, что для ошибки производного измерения величины y справедлива формула $\sigma_y \leq \Delta_y$.

7. Вывести формулы для дисперсии ошибки среднего, суммы, разности, произведения, частного от деления и возведения в степень как частные случаи общей формулы. Убедиться в том, что при суммировании и вычитании среднеквадратически складываются абсолютные ошибки, при умножении и делении - относительные ошибки.

3. Алгебра линейной регрессии

3.1. Обозначения и определения

\mathbf{x} - \mathbf{n} -вектор-строка переменных \mathbf{x}_j ;

$\boldsymbol{\alpha}$ - \mathbf{n} -вектор-столбец коэффициентов (параметров) регрессии α_j при переменных \mathbf{x} ;

β - свободный член в уравнении регрессии;

$\boldsymbol{\varepsilon}$ - ошибки измерения (ошибки уравнения, необъясненные остатки);

$\mathbf{x}\boldsymbol{\alpha} = \beta + \boldsymbol{\varepsilon}$ - уравнение (линейной) регрессии;

$\mathbf{x}\boldsymbol{\alpha} = \beta$ - гиперплоскость регрессии размерности $\mathbf{n}-1$;

$\alpha, \beta, \boldsymbol{\varepsilon}$ - истинные значения соответствующих величин;

$\mathbf{a}, \mathbf{b}, \mathbf{e}$ - их оценки;

\mathbf{x}_{-j} - вектор \mathbf{x} без j -й компоненты;

$-\alpha_{-j}$ - вектор $\boldsymbol{\alpha}$ без j -й компоненты;

\mathbf{X}_j - \mathbf{N} - вектор-столбец наблюдений $\{\mathbf{x}_{ij}\}$ за переменной \mathbf{x}_j (вектор фактических значений переменной);

\mathbf{X} - $\mathbf{N} \times \mathbf{n}$ -матрица наблюдений $\{\mathbf{X}_j\}$ за переменными \mathbf{x} ;

\mathbf{X}_{-j} - та же матрица без j -го столбца;

$\boldsymbol{\varepsilon}$ - \mathbf{N} - вектор-столбец ошибок (остатков) по наблюдениям;

$\mathbf{X}\boldsymbol{\alpha} = \mathbf{1}_N \beta + \boldsymbol{\varepsilon}$ - регрессия по наблюдениям (уравнение регрессии);

$\bar{\mathbf{x}} = \frac{1}{N} \mathbf{1}_N' \mathbf{X}$ - \mathbf{n} -вектор-строка средних;

$\bar{\mathbf{x}}_{-j}$ - тот же вектор без j -й компоненты;

$\hat{\mathbf{X}} = \mathbf{X} - \mathbf{1}_N \bar{\mathbf{x}}$ - матрица центрированных наблюдений;

$\mathbf{M} = \frac{1}{N} \hat{\mathbf{X}}' \hat{\mathbf{X}}$ - $\mathbf{n} \times \mathbf{n}$ -матрица $\{\mathbf{m}_{ij}\}$ оценок ковариаций переменных \mathbf{x} (эта

матрица, по определению, - вещественная, симметрическая и положительно полуопределенная);

\mathbf{M}_{-j} - та же матрица без j -го столбца и j -й строки;

\mathbf{m}_{-j} - $(\mathbf{n}-1)$ -вектор-столбец (оценок) ковариаций \mathbf{x}_j с остальными переменными.

$s_e^2 = \frac{1}{N} \mathbf{e}' \mathbf{e} = \frac{1}{N} (\mathbf{X}\mathbf{a} - \mathbf{1}_N \mathbf{b})' (\mathbf{X}\mathbf{a} - \mathbf{1}_N \mathbf{b})$ - оценка остаточной дисперсии.

Коэффициенты регрессии \mathbf{a} и \mathbf{b} находятся так, чтобы s_e^2 достигала своего наименьшего значения. В этом заключается применение метода наименьших квадратов.

Из условия $\frac{\partial s_e^2}{\partial \mathbf{b}} = \mathbf{0}$ определяется, что $\bar{\mathbf{e}} = \frac{1}{N} \mathbf{1}'_N \mathbf{e} = \mathbf{0}$ и $\bar{\mathbf{x}} \mathbf{a} = \mathbf{b}$, т.е.

гиперплоскость регрессии проходит через точку средних значений переменных, и ее уравнение можно записать в сокращенной форме:

$$\hat{\mathbf{X}} \mathbf{a} = \mathbf{e}.$$

3.2. Простая регрессия

Когда на вектор параметров регрессии $\boldsymbol{\alpha}$ накладывается ограничение $\boldsymbol{\alpha}_j = 1$, имеется в виду **простая регрессия**, в левой части уравнения которой остается только одна переменная:

$$\hat{\mathbf{X}}_j = \hat{\mathbf{X}}_{-j} \mathbf{a}_{-j} + \mathbf{e}$$

Это уравнение регрессии \mathbf{x}_j по \mathbf{x}_{-j} ; переменная \mathbf{x}_j - объясняемая, изучаемая или моделируемая, переменные \mathbf{x}_{-j} - объясняющие, независимые факторы, регрессоры.

Из условия $\frac{\partial s_e^2}{\partial \mathbf{a}_{-j}} = \mathbf{0}$ определяется, что $\text{cov}(\mathbf{e}, \mathbf{X}_{-j}) = \mathbf{0}$ и $\mathbf{m}_{-j} = \mathbf{M}_{-j} \mathbf{a}_{-j}$.

Последнее называется системой нормальных уравнений, из которой находятся искомые МНК-оценки параметров регрессии:

$$\mathbf{a}_{-j} = \mathbf{M}_{-j}^{-1} \mathbf{m}_{-j}.$$

Систему нормальных уравнений можно вывести, используя иную логику. Если обе части уравнения регрессии (записанного по наблюдениям) умножить слева на $\hat{\mathbf{X}}_{-j}'$ и разделить на N , то получится условие $\mathbf{m}_{-j} = \mathbf{M}_{-j} \mathbf{a}_{-j} + \frac{1}{N} \hat{\mathbf{X}}_{-j}' \mathbf{e}$, из которого следует искомая система при требованиях $\bar{\mathbf{e}} = \mathbf{0}$ и $\text{cov}(\mathbf{e}, \mathbf{X}_{-j}) = \mathbf{0}$.

Такая же логика используется в методе инструментальных переменных. Пусть имеется $N \times (n-1)$ -матрица наблюдений \mathbf{Z} за некоторыми величинами \mathbf{z} , называемыми инструментальными переменными, относительно которых известно, что они взаимно независимы с \mathbf{e} . Умножение обеих частей уравнения регрессии слева на $\hat{\mathbf{Z}}'$ и деление их на N дает условие $\frac{1}{N} \hat{\mathbf{Z}}' \hat{\mathbf{X}}_j = \frac{1}{N} \hat{\mathbf{Z}}' \hat{\mathbf{X}}_{-j} + \frac{1}{N} \hat{\mathbf{Z}}' \mathbf{e}$, из которого - после отбрасывания 2-го члена правой части - следует система нормальных уравнений

$$\mathbf{m}_{-j}^z = \mathbf{M}_{-j}^z \mathbf{a}_{-j}^z$$

метода инструментальных переменных,

где $\mathbf{m}_{-j}^z = \text{cov}(\mathbf{z}, \mathbf{x}_{-j})$, $\mathbf{M}_{-j}^z = \text{cov}(\mathbf{z}, \mathbf{x}_{-j})$.

МНК-оценка остаточной дисперсии удовлетворяет следующим формулам:

$$s_e^2 = m_{jj} - s_q^2,$$

где $s_q^2 = \mathbf{a}_{-j}' \mathbf{M}_{-j} \mathbf{a}_{-j} = \mathbf{a}_{-j}' \mathbf{m}_{-j} = \mathbf{m}_{-j}' \mathbf{a}_{-j} = \mathbf{m}_{-j}' \mathbf{M}_{-j}^{-1} \mathbf{m}_{-j}$ - объясненная дисперсия.

$$\mathbf{R}^2 = \frac{s_q^2}{\mathbf{m}_{jj}} = 1 - \frac{s_e^2}{\mathbf{m}_{jj}} \quad \text{или} \quad \frac{s_q^2}{s_j^2} = 1 - \frac{s_e^2}{s_j^2} \quad (\text{т.к. } \mathbf{m}_{jj} = s_j^2) \quad - \text{коэффициент}$$

детерминации (равный квадрату коэффициента множественной корреляции между \mathbf{x}_j и \mathbf{x}_{-j}), показывающий долю исходной дисперсии моделируемой переменной, которая объяснена регрессионной моделью.

$\hat{\mathbf{x}}_j^c = \hat{\mathbf{x}}_{-j} \mathbf{a}_{-j}$ - расчетные значения моделируемой переменной (лежащие на гиперплоскости регрессии).

В \mathbf{n} -пространстве переменных вектора-строки матрицы \mathbf{X} образуют так называемое облако наблюдений. Искомая гиперплоскость регрессии в этом пространстве располагается так, чтобы сумма квадратов расстояний от всех точек облака наблюдений до этой гиперплоскости была минимальна. Данные расстояния измеряются параллельно оси моделируемой переменной \mathbf{x}_j .

В \mathbf{N} -пространстве наблюдений показываются вектора-столбцы матрицы $\hat{\mathbf{X}}$. Коэффициент множественной корреляции между \mathbf{x}_j и \mathbf{x}_{-j} равен косинусу угла между $\hat{\mathbf{x}}_j$ и гиперплоскостью, "натянутой" на столбцы матрицы $\hat{\mathbf{X}}_{-j}$, вектор \mathbf{e} является нормалью из $\hat{\mathbf{x}}_j$ на эту гиперплоскость, а вектор \mathbf{a}_{-j} образован коэффициентами разложения проекции $\hat{\mathbf{x}}_j$ на эту гиперплоскость по векторам-столбцам матрицы $\hat{\mathbf{X}}_{-j}$.

В зависимости от того, какая переменная остается в левой части уравнения регрессии, получаются различные оценки вектора $\boldsymbol{\alpha}$ (и, соответственно, коэффициента $\boldsymbol{\beta}$). Пусть $\mathbf{a}(j)$ - оценка этого вектора из регрессии \mathbf{x}_j по \mathbf{x}_{-j} . Равенство

$$\frac{1}{\mathbf{a}_{j'}(j)} \mathbf{a}(j) = \mathbf{a}(j')$$

при $j' \neq j$ выполняется в том и только в том случае, если $\mathbf{e} = \mathbf{0}$ и, соответственно, $\mathbf{R}^2 = 1$.

При $\mathbf{n} = 2$ регрессия \mathbf{x}_1 по \mathbf{x}_2 называется прямой, регрессия \mathbf{x}_2 по \mathbf{x}_1 - обратной.

Замечание: в отечественной литературе простой обычно называют регрессию с одной переменной в правой части, а регрессию с несколькими независимыми факторами - множественной.

3.3. Ортогональная регрессия

В случае, когда ограничения на параметры α состоят в требовании равенства единице длины этого вектора

$$\alpha' \alpha = 1,$$

получается **ортогональная регрессия**, в которой расстояния от точек облака наблюдений до гиперплоскости регрессии измеряются перпендикулярно этой гиперплоскости.

Уравнение ортогональной регрессии имеет вид:

$$\hat{X}a = e, a' a = 1.$$

Теперь применение МНК означает минимизацию s_e^2 по a при указанном ограничении на длину этого вектора. Из условия равенства нулю производной по a соответствующей функции Лагранжа следует, что

$$(M - \lambda I_n)a = 0 \text{ причем } \lambda = s_e^2,$$

(λ - половина множителя Лагранжа указанного ограничения) т.е. применение МНК сводится к поиску минимального собственного числа λ ковариационной матрицы M и соответствующего ему собственного (правого) вектора a . Благодаря свойствам данной матрицы, искомые величины существуют, они вещественны, а собственное число неотрицательно (предполагается, что оно единственно). Пусть эти оценки получены.

В ортогональной регрессии все переменные x выступают изучаемыми или моделируемыми, их расчетные значения определяются по формуле

$$\hat{X}^c = \hat{X} - ea',$$

а аналогом коэффициента детерминации выступает величина

$$1 - \frac{\lambda}{s_\Sigma^2},$$

где $s_\Sigma^2 = \sum_{j=1}^n s_j^2$ - суммарная дисперсия переменных x , равная следу матрицы

M .

Таким образом, к n оценкам вектора a простой регрессии добавляется оценка этого вектора ортогональной регрессии, и общее количество этих оценок становится равным $n+1$.

Задачу простой и ортогональной регрессии можно записать в единой, обобщенной форме:

$$(M - \lambda W)a = 0, a' Wa = 1, \lambda \rightarrow \min,$$

где W - диагональная $n \times n$ -матрица, на диагонали которой могут стоять 0 или 1 .

В случае, если в матрице W имеется единственный ненулевой элемент $w_{jj} = 1$, это - задача простой регрессии x_j по x^j ; если W является единичной матрицей, то

это - задача ортогональной регрессии. Очевидно, что возможны и все промежуточные случаи, и общее количество оценок регрессии - $2^n - 1$.

Задача ортогональной регрессии легко обобщается на случай нескольких уравнений и альтернативного представления расчетных значений изучаемых переменных.

Матрица **M**, являясь вещественной, симметрической и положительно полуопределенной, имеет **n** вещественных неотрицательных собственных чисел, сумма которых равна s_{Σ}^2 , и **n** соответствующих им вещественных взаимноортогональных собственных векторов, дающих ортонормированный базис в пространстве наблюдений. Пусть собственные числа, упорядоченные по возрастанию, образуют диагональную матрицу **Λ**, а соответствующие им собственные вектора (столбцы) - матрицу **A**. Тогда

$$A^T A = I_n, \quad MA = A\Lambda.$$

Собственные вектора, если их рассматривать по убыванию соответствующих им собственных чисел, есть **главные компоненты** облака наблюдений, которые показывают направления наибольшей “вытянутости” (наибольшей дисперсии) этого облака. Количественную оценку степени этой “вытянутости” (дисперсии) дают соответствующие им собственные числа.

Пусть первые **k** собственных чисел “малы”.

s_E^2 - сумма этих собственных чисел;

A^E - часть матрицы **A**, соответствующая им (ее первые **k** столбцов); это - коэффициенты по **k** уравнениям регрессии или **k** младших главных компонент;

A^F - оставшаяся часть матрицы **A**, это - **n-k** старших главных компонент или собственно главных компонент;

$$A = [A^E, A^F];$$

$x A^E = 0$ - гиперплоскость ортогональной регрессии размерности **n-k**;

$[E, F] = \hat{X} [A^E, A^F]^T$ - координаты облака наблюдений в базисе главных компонент;

E - **N×k**-матрица остатков по уравнениям регрессии;

F - **N×(n-k)**-матрица, столбцы которой есть так называемые **главные факторы**.

Поскольку $A^T = A^{-1}$ и $AA^T = I_n$, можно записать

$$\hat{X} = [E, F] [A^E, A^F]^T = EA^{E^T} + FA^{F^T}.$$

Откуда получается два возможных представления расчетных значений переменных:

$$\hat{X}^{(1)} = \hat{X} - EA^{E^T} = FA^{F^T}.$$

Первое из них - по уравнениям ортогональной регрессии, второе (альтернативное) - по главным факторам.

$1 - \frac{s_E^2}{s_\Sigma^2}$ - аналог коэффициента детерминации, дающий оценку “качества”

этих обеих моделей.

3.4. Многообразие оценок регрессии

Множество оценок регрессии не исчерпывается $2^n - 1$ отмеченными выше элементами.

D - $N' \times N$ -матрица преобразований в пространстве наблюдений ($N' \leq N$).

Преобразование в пространстве наблюдений проводится умножением слева обеих частей уравнения регрессии (записанного по наблюдениям) на эту матрицу:

$$DXa = D1_N b + De.$$

После такого преобразования - если **D** не единичная матрица - применение МНК приводит к новым оценкам регрессии (как простой, так и ортогональной), при этом параметр **b** - если $D1_N \neq 1_N$ - теряет смысл свободного члена в уравнении.

C - невырожденная $n \times n$ -матрица преобразований в пространстве переменных.

Преобразование в пространстве переменных проводится следующим образом:

$$\hat{X} C C^{-1} a = e,$$

и в результате получается новое выражение для уравнения регрессии:

$$\hat{Y} f = e,$$

где $\hat{Y} = \hat{X} C, f = C^{-1} a.$

МНК-оценки **f** и **a** количественно различаются, если **C** не единичная матрица. Однако **f** является новой оценкой, только если $Cf \neq a$. В противном случае она совпадает с исходной оценкой **a** с точностью до сделанного преобразования (представляет ту же оценку в другой метрике или шкале измерения).

Результаты преобразования в пространстве переменных различны для простой и ортогональной регрессии.

В случае простой регрессии x_j по x_{-j} это преобразование не приводит к получению новых оценок, если **j**-я строка матрицы **C** является ортом, т.е. в независимые факторы правой части не “попадает” - после преобразования - моделируемая переменная. Если **C** диагональная матрица с элементами $c_{jj}=1$,

$c_{ii} = \frac{s_j}{s_i}$ при $i \neq j$, то оценка **f** дается в так называемой стандартизированной шкале.

Если **j**-я строка матрицы **C** имеет ненулевые внедиагональные элементы, Cf и **a** совпадают только при $R^2 = 1$.

В случае ортогональной регрессии задача определения \mathbf{f} записывается следующим образом:

$$(\mathbf{M}_Y - \lambda \mathbf{I}_n) \mathbf{f} = \mathbf{0}, \mathbf{f}' \mathbf{f} = 1,$$

где $\mathbf{M}_Y = \mathbf{C}' \mathbf{M} \mathbf{C}$.

После обратной подстановки переменных и элементарного преобразования она приобретает следующий вид:

$$(\mathbf{M} - \lambda \mathbf{\Omega}) \mathbf{a} = \mathbf{0}, \mathbf{a}' \mathbf{\Omega} \mathbf{a} = 1,$$

где $\mathbf{\Omega} = \mathbf{C}'^{-1} \mathbf{C}^{-1}$.

Решение этой задачи дает новую оценку, даже если \mathbf{C} является диагональной матрицей. Это - так называемая **регрессия в метрике $\mathbf{\Omega}^{-1}$** .

Теоретические вопросы и задания

1. Доказать, что матрица ковариации является симметрической и положительно полуопределенной. В каком случае она положительно определена?

2(**). Показать, что гиперплоскость регрессии проходит через точку средних значений переменных, и оценки остатков имеют нулевую среднюю.

3(**). Вывести систему нормальных уравнений для коэффициентов при независимых факторах простой регрессии.

4(*). Доказать, что оценки остатков в простой регрессии не скоррелированы с независимыми факторами.

5(*). Вывести формулу для остаточной дисперсии в простой регрессии.

6(*). Провести геометрическую иллюстрацию простой регрессии в пространстве переменных и наблюдений, убедиться в справедливости сделанных выше утверждений относительно геометрических образов объектов регрессии.

7. Доказать, что оценки параметров прямой и обратной регрессии совпадают в случае и только в случае функциональной зависимости между переменными.

8(**). Показать, что МНК в ортогональной регрессии сводится к поиску собственных чисел и векторов ковариационной матрицы. Почему остаточная дисперсия равна минимальному собственному числу этой матрицы?

9(*). Почему для определения расчетных значений переменных в ортогональной регрессии используется приведенная формула?

10(*). Дать геометрическую иллюстрацию ортогональной регрессии, главным компонентам и главным факторам в пространстве переменных.

11. В каком случае преобразование в пространстве наблюдений можно применять к сокращенной форме уравнения регрессии? Почему преобразование в пространстве переменных всегда применимо к сокращенной форме уравнения?

12(*). Доказать, что в случае простой регрессии преобразование в пространстве переменных приводит к новым оценкам только в случае, если независимая переменная в результате проведенного преобразования “попадает” в правую часть уравнения. Показать, что в таком случае оценки все-таки не меняются, если зависимость между переменными функциональная.

13(*). Показать, что оценки простой регрессии в стандартизированной шкале получаются, если в системе нормальных уравнений использовать не ковариационную, а корреляционную матрицу.

14. Вывести приведенную формулу для оценки регрессии в метрике $\mathbf{\Omega}^{-1}$.

15(*). Совпадают ли полученные по ковариационной и корреляционной матрице оценки ортогональной регрессии и главных компонент с точностью до обратного преобразования?

4. Основная модель линейной регрессии

4.1. Различные формы уравнения регрессии

X - моделируемая переменная;

Z - n -вектор-строка независимых факторов;

$x = Z\alpha + \beta + \epsilon$ - уравнение регрессии;

X , Z - N -вектор и $N \times n$ -матрица наблюдений за соответствующими переменными;

\bar{Z} - n -вектор-строка средних значений переменных Z .

Первые две формы уравнения регрессии по наблюдениям аналогичны используемым в предыдущем разделе и имеют следующий вид:

$$X = Z\alpha + 1_N\beta + \epsilon,$$

или $X = Za + 1_N b + e$ (истинные значения заменены их оценками)

- исходная форма;

$$\hat{X} = \hat{Z}a + e$$

- сокращенная форма.

Оператор МНК-оценивания для этих двух форм имеет следующий вид:

$$a = M^{-1}m, \quad b = \bar{x} - \bar{z}a,$$

где $M = \frac{1}{N} \hat{Z}' \hat{Z}$ - $n \times n$ -матрица ковариации (вторых центральных моментов)

Z ;

$$m = \frac{1}{N} \hat{Z}' \hat{X} - n\text{-вектор-столбец ковариации между } Z \text{ и } X.$$

Третья форма - без свободного члена - записывается следующим образом:

$$X = Za + e,$$

где Z - $N \times (n+1)$ -матрица, последний столбец которой состоит из единиц (равен 1_N);

a - $(n+1)$ -вектор-столбец, последний элемент которого является свободным членом регрессии.

Какая из этих форм регрессии используется и, соответственно, что именно означают a и Z , будет в дальнейшем ясно из контекста или будет специально поясняться.

В этом разделе, в основном, используется форма уравнения регрессии без свободного члена.

Оператор МНК-оценивания для нее записывается более компактно:

$$a = M^{-1}m,$$

но $M = \frac{1}{N} Z'Z$ - $(n+1) \times (n+1)$ -матрица вторых **начальных** моментов

$[z, 1]$;

$\mathbf{m} = \frac{1}{N} \mathbf{Z}' \mathbf{X}$ - $(n+1)$ -вектор-столбец вторых **начальных** моментов между

$[\mathbf{z}, 1]$ и \mathbf{x} .

Если в этом операторе вернуться к обозначениям первых двух форм уравнения регрессии, то получится следующее выражение:

$$\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{M}^{-1} & -\mathbf{M}^{-1} \bar{\mathbf{z}}' \\ -\bar{\mathbf{z}} \mathbf{M}^{-1} & 1 + \bar{\mathbf{z}} \mathbf{M}^{-1} \bar{\mathbf{z}}' \end{bmatrix} \begin{bmatrix} \mathbf{m} + \bar{\mathbf{z}}' \bar{\mathbf{x}} \\ \bar{\mathbf{x}} \end{bmatrix},$$

из которого видно, что

- обратная матрица ковариации \mathbf{Z} (размерности $N \times N$) совпадает с соответствующим блоком обратной матрицы вторых начальных моментов (размерности $(N+1) \times (N+1)$);

- результаты применения двух приведенных операторов оценивания одинаковы.

4.2. Основные гипотезы, свойства оценок

1. Между переменными \mathbf{x} и \mathbf{z} существует зависимость $\mathbf{x} = \mathbf{z}\boldsymbol{\alpha} + \boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

2. Переменные \mathbf{z} детерминированы, наблюдаются без ошибок и линейно независимы (в алгебраическом смысле).

3. $\mathbf{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\mathbf{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2 \mathbf{I}_N$.

4. В модели линейной регрессии математической статистики, в которой переменные \mathbf{z} случайны, предполагается, что ошибки $\boldsymbol{\varepsilon}$ не зависят от них и - по крайней мере - не скоррелированы с ними. В данном случае это предположение формулируется так: независимо от того, какие значения принимают переменные \mathbf{z} , ошибки $\boldsymbol{\varepsilon}$ удовлетворяют гипотезе 3.

В этих предположениях \mathbf{a} относится к классу линейных оценок, т.к.

$$\mathbf{a} = \mathbf{L}\mathbf{X},$$

где $\mathbf{L} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ - неслучайный $(n+1) \times (N+1)$ -оператор оценивания;

а также доказывается что

- \mathbf{a} является несмещенной оценкой $\boldsymbol{\alpha}$, их матрица ковариации \mathbf{M}_a равна

$\frac{1}{N} \sigma^2 \mathbf{M}^{-1}$ (в обозначениях сокращенной формы уравнения регрессии это выражение давало бы - как показано в предыдущем пункте - матрицу ковариации коэффициентов регрессии при независимых факторах, а дисперсия свободного члена

определялась бы по формуле $\frac{\sigma^2}{N} (1 + \bar{\mathbf{z}} \mathbf{M}^{-1} \bar{\mathbf{z}}')$), и дисперсия любой их линейной

комбинации минимальна на множестве линейных оценок, т.е. они относятся к классу **BLUE** - **Best Linear Unbiased Estimators**;

- несмещенной оценкой σ^2 является

$$\hat{s}_e^2 = \frac{N}{N - n - 1} s_e^2 = \frac{1}{N - n - 1} \mathbf{e}' \mathbf{e}.$$

Для расчета коэффициента детерминации можно использовать следующую формулу:

$$R^2 = \frac{q - \bar{x}^2}{m_{xx} - \bar{x}^2},$$

где $q = a' M a = m' M^{-1} m = m' a = a' m$,

$$m_{xx} = \frac{1}{N} X' X.$$

Если предположить, что ϵ (и, следовательно, их оценки e) распределены нормально:

$$\epsilon \sim N(0, \sigma^2 I_N),$$

то оценки a также будут иметь нормальное распределение:

$$a \sim N(\alpha, M_a),$$

они совпадут с оценками максимального правдоподобия, будут несмещенными, состоятельными и эффективными.

В этом случае можно строить доверительные интервалы для оценок и использовать статистические критерии проверки гипотез.

(1-θ)100-процентный доверительный интервал для α_i , $i = 1, \dots, n+1$ ($\alpha_{n+1} = \beta$), строится следующим образом:

$$a_i \pm s_{a_i} t_{N-n-1, 1-\theta},$$

где $s_{a_i} = \hat{s}_e \sqrt{\frac{1}{N} m_{ii}^{-1}}$ - среднеквадратическое отклонение a_i (m_{ii}^{-1} - ii-й

элемент матрицы M^{-1});

$t_{N-n-1, 1-\theta}$ - **(1-θ)100-процентный** двусторонний квантиль t_{N-n-1} -распределения.

Для проверки нулевой гипотезы $\alpha_i = 0$ применяется **t-критерий**. Гипотеза отвергается (влияние i-го фактора считается статистически значимым) с вероятностью ошибки (1-го рода) θ , если

$$\frac{|a_i|}{s_{a_i}} \geq t_{N-n-1, 1-\theta},$$

т.к. при выполнении нулевой гипотезы величина $\frac{a_i}{s_{a_i}}$ имеет t_{N-n-1} -распределение.

Эта величина называется **t-статистикой** (**t_i-статистикой**) и ее фактическое значение обозначается в дальнейшем **t_i^c**.

При использовании современных статистических пакетов программ не требуется искать нужные квантили **t**-распределения в статистических таблицах, поскольку в них (пакетах) рассчитывается уровень ошибки θ_i^c , с которой можно отвергнуть нулевую гипотезу, т.е. такой, что:

$$|t_i^c| = t_{N-n-1, 1-\theta_i^c},$$

и, если он меньше желаемого значения либо равен ему, то нулевая гипотеза отвергается.

Для проверки нулевой гипотезы об отсутствии искомой связи $\alpha_i = 0, i = 1, \dots, n$ применяется **F-критерий**. Если эта гипотеза верна, величина

$$\frac{R^2(N - n - 1)}{(1 - R^2)n}$$

имеет $F_{n,N-n-1}$ -распределение. Данная величина называется **F-статистикой** и ее фактическое значение обозначается в дальнейшем F^c . Нулевая гипотеза отвергается (влияние **Z** на **X** считается статистически значимым) с вероятностью ошибки (1-го рода) θ , если

$$F^c \geq F_{n,N-n-1,1-\theta},$$

где $F_{n,N-n-1,1-\theta}$ - $(1-\theta)100$ -процентный (односторонний) квантиль $F_{n,N-n-1}$ -распределения.

В современных статистических пакетах программ также рассчитывается уровень θ^c ошибки для F^c , такой, что

$$F^c = F_{n,N-n-1,1-\theta^c}.$$

Уместно отметить, что приведенные в разделе 2.1. сведения являются частным случаем рассмотренных здесь результатов при $n=0$.

4.3. Независимые факторы

Если не выполняется 2-я гипотеза, и некоторые из переменных **Z** линейно зависят от других, то матрица **M** вырождена, и использование приведенного оператора оценивания невозможно.

Вообще говоря, предложить метод оценивания параметров регрессии в этом случае можно. Так, пусть множество независимых факторов разбито на две части (в этом фрагменте используются обозначения сокращенной формы уравнения регрессии):

$$z = [z_1, z_2], \hat{Z} = [\hat{Z}_1, \hat{Z}_2], a = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix},$$

и $\hat{Z}_2 = \hat{Z}_1 C_{12}.$

Тогда можно записать уравнение регрессии в форме

$$\hat{X} = \hat{Z}_1(a_1 + C_{12}a_2) + e,$$

и оценить линейную комбинацию параметров $a_1 + C_{12}a_2$ (предполагая, что столбцы **Z**₁ линейно независимы). Но чтобы оценить сами параметры, нужна априорная информация, например: $a_2 = 0$.

Однако вводить в регрессию факторы, которые линейно зависят от уже введенных факторов, не имеет смысла, т.к. при этом не растет объясненная дисперсия (см. ниже).

На практике редко встречается ситуация, когда матрица **M** вырождена. Более распространен случай, когда она плохо обусловлена (между переменными **Z** существуют зависимости близкие к линейным). В этом случае имеет место **мультиколлинеарность факторов**. Поскольку гипотеза 2 в части отсутствия ошибок измерения, как правило, нарушается, получаемые (при мультиколлинеарности) оценки в значительной степени обусловлены этими ошибками измерения. В таком случае (если связь существует), обычно, факторы по отдельности оказываются незначимыми по **t**-критерию, а все вместе - существенными по **F**-критерию. Поэтому в регрессию стараются не вводить факторы сильно скоррелированные с остальными.

В общем случае доказывается, что

$$s_q^2 = s_{q1}^2 + \Delta s_{q12}^2 \leq s_x^2, \quad 0 \leq \Delta s_{q12}^2 \leq s_{q2}^2,$$

где s_{q1}^2 и s_{q2}^2 - дисперсии, объясненные факторами **z₁** и **z₂** по отдельности;

Δs_{q12}^2 - прирост объясненной дисперсии, вызванный добавлением в регрессии факторов **z₂** к факторам **z₁**.

В соотношении для прироста объясненной дисперсии:

- левая часть выполняется как строгое равенство, если и только если

$s_{q1}^2 = s_x^2$ (коэффициент детерминации в регрессии по **z₁** уже равен единице), или

вектор остатков в регрессии по **z₁** ортогонален факторам $\hat{\mathbf{z}}_2$, т.е. имеет с ними нулевую корреляцию (возможное влияние факторов **z₂** уже “приняли” на себя факторы **z₁**), или

факторы $\hat{\mathbf{z}}_2$ линейно зависят от факторов $\hat{\mathbf{z}}_1$;

- правая часть выполняется как строгое равенство, если и только если

факторы $\hat{\mathbf{z}}_2$ ортогональны факторам $\hat{\mathbf{z}}_1$.

Если в множество линейно независимых факторов добавлять новые элементы, то коэффициент детерминации растет вплоть до единицы, после чего рост прекращается. Своего максимального значения он обязательно достигнет при **n** = **N** (возможно и раньше) - даже если вводимые факторы не влияют по-существу на изучаемую переменную. Поэтому сам по себе коэффициент детерминации не может служить статистическим критерием “качества” уравнения регрессии. Более приемлем в этой роли коэффициент детерминации, скорректированный на число степеней свободы:

$$\overline{R^2} = 1 - (1 - R^2) \frac{N - 1}{N - n - 1},$$

который может и уменьшиться при введении нового фактора. Точную же статистическую оценку качества (в случае нормальности распределения остатков) дает **F**-критерий. Однако учитывая, что значения **F^c** оказываются несопоставимыми при изменении **n** (т.к. получают разное число степеней свободы), наиболее правильно эту роль возложить на уровень ошибки **θ^c** для **F^c**.

В результате введения новых факторов в общем случае меняются оценки параметров при ранее введенных факторах:

$$\mathbf{a}_1 = \mathbf{a}_1^0 + \mathbf{A}_{12}\mathbf{a}_2,$$

где \mathbf{a}_1^0 - оценка параметров регрессии по \mathbf{z}_1 (до введения новых факторов);

\mathbf{A}_{12} - матрица, столбцы которой являются оценками параметров регрессии переменных \mathbf{z}_2 по \mathbf{z}_1 .

“Старые” оценки параметров сохраняются ($\mathbf{a}_1 = \mathbf{a}_1^0$), если и только если

- коэффициент детерминации в регрессии по \mathbf{z}_1 уже равен единице, или

вектор остатков в регрессии по \mathbf{z}_1 ортогонален факторам $\hat{\mathbf{Z}}_2$ (в этих двух случаях $\mathbf{a}_2 = \mathbf{0}$), или

факторы $\hat{\mathbf{Z}}_2$ ортогональны факторам $\hat{\mathbf{Z}}_1$ (в этом случае $\mathbf{A}_{12} = \mathbf{0}$).

Итак, возникает проблема определения истинного набора факторов, фигурирующих в гипотезе 1, который позволил бы найти оценки истинных параметров регрессии. Определение такого набора факторов есть **спецификация модели**. Формальный подход к решению этой проблемы заключается в поиске так называемого наилучшего уравнения регрессии, для чего используется **процесс (метод) шаговой регрессии**.

Пусть \mathbf{z} - полный набор факторов, потенциально влияющих на \mathbf{x} . Рассматривается процесс обращения матрицы ковариации переменных $[\mathbf{x}, \mathbf{z}]$. В паре матриц $(\mathbf{n}+1) \times (\mathbf{n}+1)$

$$\begin{bmatrix} \mathbf{m}_{xx} & \mathbf{m}_{x1} & \mathbf{m}_{x2} & . & . & \mathbf{m}_{xn} \\ \mathbf{m}_{1x} & \mathbf{m}_{11} & \mathbf{m}_{12} & . & . & \mathbf{m}_{1n} \\ \mathbf{m}_{2x} & \mathbf{m}_{21} & \mathbf{m}_{22} & . & . & \mathbf{m}_{2n} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \mathbf{m}_{nx} & \mathbf{m}_{n1} & \mathbf{m}_{n2} & . & . & \mathbf{m}_{nn} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & . & . & 0 \\ 0 & 1 & 0 & . & . & 0 \\ 0 & 0 & 1 & . & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & 1 \end{bmatrix}$$

делаются одновременные преобразования их строк в орты. Известно, что, если 1-ю матрицу преобразовать в единичную, то на месте 2-й матрицы будет получена обратная к 1-й (исходной). Пусть этот процесс не завершен, и только несколько строк 1-й матрицы (но не ее 1-я строка) преобразованы в орты. Это - ситуация на текущем шаге процесса.

На этом шаге строкам-ортам в 1-й матрице соответствуют включенные в регрессию факторы, на их месте в 1-й строке этой матрицы оказываются текущие оценки параметров регрессии при них. Строкам-ортам во 2-й матрице соответствуют невведенные факторы, на их месте в 1-й строке 1-й матрицы размещаются коэффициенты ковариации этих факторов с текущими остатками изучаемой переменной. На месте \mathbf{m}_{xx} показывается текущее значение остаточной дисперсии.

На каждом шаге оцениваются последствия введения в регрессию каждого не включенного фактора (преобразованием в орты соответствующих строк 1-й матрицы) и исключения каждого введенного ранее фактора (преобразованием в орты соответствующих строк 2-й матрицы). Выбирается тот вариант, который дает

минимальный уровень ошибки θ^c для F^c . Процесс продолжается до тех пор, пока этот уровень сокращается.

Иногда в этом процессе используются более простые критерии. Например, задается определенный уровень t -статистики (правильнее - уровень ошибки θ^c для t^c), и фактор вводится в уравнение, если фактическое значение t^c для него выше заданного уровня (ошибка θ^c ниже ее заданного уровня), фактор исключается из уравнения в противном случае.

Такие процессы, как правило, исключают возможность введения в уравнение сильно скоррелированных факторов, т.е. решают проблему мультиколлинеарности.

Формальные подходы к спецификации модели должны сочетаться с теоретическими подходами, когда набор факторов и, часто, знаки параметров регрессии определяются из теории изучаемого явления.

4.4. Прогнозирование

Требуется определить наиболее приемлемое значения для x_{N+1} (прогноз), если известны значения независимых факторов (вектор-строка):

$$z_{N+1} = [z_{1,N+1}, \dots, z_{n,N+1}, 1].$$

$x_{N+1} = z_{N+1}\alpha + \varepsilon_{N+1}$ - истинное значение искомой величины;

$x_{N+1}^0 = E(x_{N+1}) = z_{N+1}\alpha$ - ожидаемое значение;

$x_{N+1}^p = z_{N+1}a$ - искомый МНК-прогноз.

Полученный прогноз не смещен относительно ожидаемого значения:

$$E(x_{N+1}^p) = x_{N+1}^0,$$

и его ошибка $d = x_{N+1} - x_{N+1}^p$ имеет нулевое матожидание:

$$E(d) = 0,$$

и дисперсию $\sigma_d^2 = \sigma^2(1 + \frac{1}{N} z_{N+1} M^{-1} z_{N+1}')$, которая минимальна в классе линейных оценок α .

Оценка стандартной ошибки прогноза при $n = 1$ рассчитывается по формуле

$$\hat{s}_e^2 \sqrt{1 + \frac{1}{N} + \frac{(z_{N+1} - \bar{z})^2}{\sum_{i=1}^N (z_i - \bar{z})^2}}.$$

Теоретические вопросы и задания

1. Провести матричные преобразования, доказывающие эквивалентность операторов оценивания для первых двух (основная и сокращенная) и третьей (без свободного члена) форм уравнения регрессии.

2(*). Показать, что $e = B\varepsilon$,

где $\mathbf{B} = \mathbf{I} - \frac{1}{N} \mathbf{Z} \mathbf{M}^{-1} \mathbf{Z}'$ - симметрическая, идемпотентная и положительно

полуопределенная матрица.

3(**). Доказать принадлежность МНК-оценок регрессии классу BLUE.

4(**). Вывести приведенную формулу для матрицы \mathbf{M}_a ковариации оценок.

5(**). Показать, что \hat{s}_a^2 является несмещенной оценкой дисперсии ошибок

σ^2 .

6. Вывести приведенную формулу для расчета коэффициента детерминации.

7(*). Доказать, что при нормальности распределения остатков $\boldsymbol{\varepsilon}$ МНК-оценки регрессии совпадают с оценками максимального правдоподобия.

8(*). Почему в случае незначимости влияния i -го фактора t_i -статистика имеет t_{N-n-1} -распределение?

9(*). Почему в случае незначимости влияния всех факторов F -статистика имеет $F_{n, N-n-1}$ -распределение?

10(*). Проверить справедливость приведенного соотношения для прироста объясненной дисперсии, вызванного введением в регрессию новых факторов. Почему это соотношение выполняется как равенство в указанных и только в указанных случаях?

11. Как получена формула для коэффициента детерминации, скорректированного на число степеней свободы?

12(*). Показать, что добавление новых факторов в регрессию не меняет “старые” оценки параметров в указанных и только в указанных случаях.

13(*). Убедиться в справедливости сделанных утверждений о характере заполнения указанных матриц на текущем шаге процесса шаговой регрессии.

14(*). Вывести приведенную формулу дисперсии ошибки прогноза.

15(*). Доказать указанные свойства ошибки прогноза.

16(*). Вывести приведенную формулу для оценки стандартной ошибки прогноза при $n = 1$, объяснить составляющие этой ошибки.

5. Гетероскедастичность и автокорреляция ошибок

5.1. Обобщенный метод наименьших квадратов (взвешенная регрессия)

Если матрица ковариации ошибок по наблюдениям отлична от $\sigma^2 \mathbf{I}_N$ (нарушена 3-я гипотеза основной модели), то МНК-оценки параметров регрессии остаются несмещенными, но перестают быть эффективными в классе линейных. Смещенными оказываются МНК-оценки их ковариации, в частности оценки их стандартных ошибок (как правило, они преуменьшаются).

Пусть теперь $\mathbf{E}(\mathbf{e}\mathbf{e}') = \sigma^2 \mathbf{\Omega}$, где $\mathbf{\Omega}$ - вещественная, симметрическая положительно определенная матрица (структура ковариации ошибок). **Обобщенный метод наименьших квадратов (ОМНК)**, приводящий к оценкам класса **BLUE**, означает минимизацию взвешенной суммы квадратов отклонений:

$$\frac{1}{N} \mathbf{e}' \mathbf{\Omega}^{-1} \mathbf{e}.$$

Для доказательства проводится преобразование в пространстве наблюдений с помощью невырожденной $N \times N$ -матрицы \mathbf{D} , такой, что $\mathbf{D}^{-1} \mathbf{D}'^{-1} = \mathbf{\Omega}$. После такого преобразования остатки $\mathbf{D}\mathbf{e}$ начинают удовлетворять 2-й гипотезе.

На практике с матрицами $\mathbf{\Omega}$ общего вида обычно не работают. Рассматривается два частных случая.

5.2. Гетероскедастичность ошибок

Пусть ошибки не скоррелированы по наблюдениям, и матрица $\mathbf{\Omega}$ диагональна. Если эта матрица единична, т.е. дисперсии ошибок одинаковы по наблюдениям (гипотеза 3 не нарушена), то имеет место **гомоскедастичность** или однородность ошибок по дисперсии. В противном случае констатируют **гетероскедастичность ошибок** или их неоднородность по дисперсии.

Для проверки гипотезы о гомоскедастичности можно использовать **критерий Бартлета**. Для расчета \mathbf{b}^c - статистики, лежащей в основе применения этого критерия, множество МНК-оценок остатков \mathbf{e}_i , $i = 1, \dots, N$ делится на k непересекающихся подмножеств.

N_l - количество элементов в l -м подмножестве, $\sum_{l=1}^k N_l = N$;

s_l^2 - оценка дисперсии в l -м подмножестве;

$$b_s = \frac{\frac{1}{N} \sum_{l=1}^k N_l s_l^2}{\left(\prod_{l=1}^k s_l^{2N_l} \right)^{\frac{1}{N}}} - \text{отношение средней арифметической дисперсий к}$$

средней геометрической; это отношение больше или равно единице, и чем сильнее различаются дисперсии по подмножествам, тем оно выше;

$$b^c = \frac{N}{1 + \frac{\sum_{l=1}^k \frac{1}{N_l} - \frac{1}{N}}{3(k-1)}} \ln b_s.$$

При однородности наблюдений по дисперсии эта статистика распределена как χ_{k-1}^2 .

Факт неоднородности наблюдений по дисперсии остатков мало сказывается на качестве оценок регрессии, если эти дисперсии не скоррелированы с независимыми факторами. Проверить наличие зависимости дисперсии ошибок от факторов-регрессоров можно следующим образом.

Все наблюдения упорядочиваются по возрастанию одного из независимых факторов или расчетного значения изучаемой переменной **Za**. Оценивается остаточная дисперсия s_1^2 по **K** “малым” и s_2^2 по **K** “большим” наблюдениям (“средние” **N-2K** наблюдения в расчете не участвуют, а **K** выбирается приблизительно равным трети **N**). В случае гомоскедастичности ошибок отношение $\frac{s_2^2}{s_1^2}$ распределено как $F_{K-n-1, K-n-1}$.

Если гипотеза гомоскедастичности отвергается, необходимо дать оценку матрице **Ω**. Совместить проверку этой гипотезы с оценкой данной матрицы можно следующим образом.

В качестве оценок дисперсии ошибок по наблюдениям принимаются квадраты оценок остатков e_i^2 , и строится регрессия $|e_i|$ на все множество независимых факторов или какое-то их подмножество. Если какая-то из этих регрессий оказывается статистически значимой, то гипотеза гомоскедастичности отвергается, и в качестве оценок $\sqrt{\omega_{ii}}$ ($\omega_{ii'} = 0, i \neq i'$ по предположению) принимаются расчетные значения $|e_i|^c$.

В некоторых статистических критериях проверки на гомоскедастичность в качестве оценок ω_{ii} принимаются непосредственно e_i^2 .

Имея оценку матрицы **Ω**, можно провести преобразование в пространстве наблюдений с помощью матрицы $D = \Omega^{-\frac{1}{2}}$, после которого остатки **Dε** можно считать удовлетворяющими гипотезе 3.

5.3. Автокорреляция ошибок

Пусть теперь наблюдения однородны по дисперсии и их последовательность имеет физический смысл и жестко фиксирована (например, наблюдения проводятся в последовательные моменты времени).

Для проверки гипотезы о наличии линейной автокорреляции 1-го порядка ошибок по наблюдениям

$$\varepsilon_i = \rho \varepsilon_{i-1} + \eta_i, E(\eta) = 0, E(\eta\eta') = \sigma_\eta^2 I_N,$$

где ρ - коэффициент авторегрессии 1-го порядка;

η - N -вектор-столбец $\{\eta_i\}$;

можно использовать **критерий Дарбина-Уотсона** или **DW-критерий** (при автокорреляции 2-го и более высоких порядков его применение становится ненадежным).

Фактическое значение d^c статистики Дарбина-Уотсона (отношения Фон-Неймана) или **DW**-статистики рассчитывается следующим образом:

$$d^c = \frac{\sum_{i=2}^N (e_i - e_{i-1})^2}{\sum_{i=1}^N e_i^2}$$

Оно лежит в интервале от 0 до 4, в случае отсутствия автокорреляции ошибок приблизительно равно 2, при положительной автокорреляции смещается в меньшую сторону, при отрицательной - в большую сторону.

Если $\rho = 0$, величина d распределена нормально, но параметры этого распределения зависят не только от N и n . Поэтому существует по два значения для каждого (двустороннего) квантиля, соответствующего определенным θ , N и n : его нижняя d_L и верхняя d_U границы. Нулевая гипотеза принимается, если $d_U \leq d^c \leq 4 - d_U$; она отвергается в пользу гипотезы о положительной автокорреляции, если $d^c < d_L$, и в пользу гипотезы об отрицательной автокорреляции, если $d^c > 4 - d_L$. Если $d_L \leq d^c < d_U$ или $4 - d_U < d^c \leq 4 - d_L$, вопрос остается открытым (это - зона неопределенности **DW**-критерия).

Пусть нулевая гипотеза отвергнута. Тогда необходимо дать оценку матрицы Ω .

Оценка r параметра авторегрессии ρ определяется из приближенного равенства

$$r \approx 1 - \frac{d^c}{2},$$

или рассчитывается непосредственно из регрессии e на него самого со сдвигом на одно наблюдение.

Оценкой матрицы $\mathbf{\Omega}$ является
$$\begin{bmatrix} 1 & \mathbf{r} & \mathbf{r}^2 & . & . & \mathbf{r}^{N-1} \\ \mathbf{r} & 1 & \mathbf{r} & . & . & \mathbf{r}^{N-2} \\ \mathbf{r}^2 & \mathbf{r} & 1 & . & . & \mathbf{r}^{N-3} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \mathbf{r}^{N-1} & \mathbf{r}^{N-2} & \mathbf{r}^{N-3} & . & . & 1 \end{bmatrix}, \text{ а матрица } \mathbf{D}$$

преобразований в пространстве наблюдений равна
$$\begin{bmatrix} \sqrt{1-\mathbf{r}^2} & 0 & 0 & . & . & 0 \\ -\mathbf{r} & 1 & 0 & . & . & 0 \\ 0 & -\mathbf{r} & 1 & . & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & 1 \end{bmatrix}.$$

Для преобразования в пространстве наблюдений, называемом в данном случае авторегрессионным, используют обычно указанную матрицу без 1-й строки, что ведет к сокращению количества наблюдений на одно. В результате такого преобразования из каждого наблюдения, начиная со 2-го, вычитается предыдущее, умноженное на \mathbf{r} , теоретическими остатками становятся $\boldsymbol{\eta}_i$, которые удовлетворяют гипотезе 2.

После этого преобразования снова оцениваются параметры регрессии. Если новое значение **DW**-статистики неудовлетворительно, то можно провести следующее авторегрессионное преобразование.

Обобщает процедуру последовательных авторегрессионных преобразований **метод Кочрена-Оркарта**, который заключается в следующем.

Для одновременной оценки \mathbf{r} , \mathbf{a} и \mathbf{b} используется критерий ОМНК (в обозначениях исходной формы уравнения регрессии):

$$\frac{1}{N} \sum_{i=2}^N ((\mathbf{x}_i - \mathbf{r}\mathbf{x}_{i-1}) - (\mathbf{z}_i - \mathbf{r}\mathbf{z}_{i-1})\mathbf{a} - (1 - \mathbf{r})\mathbf{b})^2 \rightarrow \min,$$

где \mathbf{z}_i - \mathbf{n} -вектор-строка значений независимых факторов в i -м наблюдении (i -строка матрицы \mathbf{Z}).

Поскольку производные функционала по искомым величинам нелинейны относительно них, применяется итеративная процедура, на каждом шаге которой сначала оцениваются \mathbf{a} и \mathbf{b} при фиксированном значении \mathbf{r} предыдущего шага (на первом шаге обычно $\mathbf{r} = 0$), а затем - \mathbf{r} при полученных значениях \mathbf{a} и \mathbf{b} . Процесс, как правило, сходится.

Теоретические вопросы и задания

1(*) Почему нарушение гипотезы 3 в части матрицы ковариации ошибок сохраняет несмещенность оценок, но приводит к потере их эффективности в классе линейных оценок?

2. Построить оператор ОМНК-оценивания, вывести формулу для матрицы ковариации оценок параметров в этом случае.

3(*). Показать, что ОМНК-оценки относятся к классу **BLUE**.

4. Убедиться, что в случае гетероскедастичности ошибок для преобразования в пространстве наблюдений используется указанная матрица.

5(*). Почему при использовании критерия Дарбина-Уотсона требуется знать два критических значения для расчетной статистики?

6. Доказать, что в случае автокорреляции ошибок 1-го порядка матрица ковариации ошибок по наблюдениям и матрица авторегрессионного преобразования имеют указанную форму.

7. Вывести формулу ОМНК-критерия и построить процедуру оценивания коэффициента авторегрессии в методе Кочрена-Оркарта.

6. Ошибки измерения факторов и фиктивные переменные

6.1. Ошибки измерения факторов

Пусть теперь нарушается гипотеза 2, и независимые факторы наблюдаются с ошибками (здесь используются обозначения первых двух форм уравнения регрессии):

$$\mathbf{z} = \mathbf{z}^0 + \boldsymbol{\varepsilon}, \text{ или в разрезе наблюдений: } \mathbf{Z} = \mathbf{Z}^0 + \boldsymbol{\varepsilon},$$

где \mathbf{z}^0 и $\boldsymbol{\varepsilon}$ - \mathbf{n} -вектора-строки истинных значений факторов и ошибок их измерений;

\mathbf{Z}^0 и $\boldsymbol{\varepsilon}$ - соответствующие $\mathbf{N} \times \mathbf{n}$ -матрицы значений этих величин по наблюдениям.

Предполагается, что истинные значения и ошибки независимы друг от друга (по крайней мере, не скоррелированы друг с другом) и известны их матрицы ковариации (одинаковые для всех наблюдений):

$$\mathbf{E}(\mathbf{z}^{0'}, \boldsymbol{\varepsilon}) = \mathbf{0}, \quad \mathbf{E}(\mathbf{z}^{0'}, \mathbf{z}^0) = \mathbf{M}^0, \quad \mathbf{E}(\boldsymbol{\varepsilon}' \boldsymbol{\varepsilon}) = \boldsymbol{\Omega}.$$

Уравнение регрессии можно записать в следующей форме:

$$\hat{\mathbf{X}} = \hat{\mathbf{Z}} \boldsymbol{\alpha} + \boldsymbol{\varepsilon} - \boldsymbol{\varepsilon} \boldsymbol{\alpha},$$

(т.е. остатки теперь не могут быть независимыми от факторов-регрессоров) и в рамках сделанных предположений доказать, что

$$\mathbf{E}(\mathbf{M}) = \mathbf{M}^0 + \boldsymbol{\Omega}, \quad \mathbf{E}(\mathbf{a}) = (\mathbf{M}^0 + \boldsymbol{\Omega})^{-1} \mathbf{M}^0 \boldsymbol{\alpha},$$

т.е. МНК-оценки теряют в такой ситуации даже свойство несмещенности. Как правило, они преуменьшены по сравнению с истинными значениями (в случае $\mathbf{n} = \mathbf{1}$,

$$\mathbf{E}(\mathbf{a}) = \frac{\sigma_{z^0}^2}{\sigma_{z^0}^2 + \sigma_{\varepsilon}^2} \boldsymbol{\alpha}).$$

Существуют три подхода к оценке параметров регрессии в случае наличия ошибок измерения независимых факторов.

а) Простая регрессия. Если имеется оценка \mathbf{W} ковариационной матрицы ошибок $\boldsymbol{\Omega}$ и ошибки регрессоров взаимно независимы с изучаемой переменной, то можно использовать следующий оператор оценивания:

$$\mathbf{a} = (\mathbf{M} - \mathbf{W})^{-1} \mathbf{m},$$

который обеспечивает несмещенность оценок.

б) Инструментальные переменные. Если имеется \mathbf{n} факторов \mathbf{y} , которые взаимно независимы как с ошибками уравнения $\boldsymbol{\varepsilon}$, так и ошибками основных факторов $\boldsymbol{\varepsilon}$, то оценка

$$\mathbf{a} = (\hat{\mathbf{Y}}' \hat{\mathbf{Z}})^{-1} (\hat{\mathbf{Y}}' \hat{\mathbf{X}})$$

несмещена.

Исторически первой в этом классе получена **оценка Вальда** для случая $\mathbf{n} = \mathbf{1}$. Для получения этой оценки i -я компонента вектора-столбца \mathbf{Y} принимается равной единице, если \mathbf{z}_i больше своей медианы, и минус единице, если - меньше медианы (при нечетном \mathbf{N} среднее значение теряется). В результате получается, что

$$\mathbf{a} = \frac{\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1}{\bar{\mathbf{z}}_2 - \bar{\mathbf{z}}_1}$$

где $\bar{\mathbf{x}}_2, \bar{\mathbf{z}}_2$ - средние значения переменных по верхней части выборки,
 $\bar{\mathbf{x}}_1, \bar{\mathbf{z}}_1$ - их средние значения по нижней части выборки.

Такая оценка более эффективна, если исключить примерно треть “средних” наблюдений.

Позже эта оценка была обобщена: матрицу значений инструментальных переменных было предложено формировать столбцами рангов по наблюдениям соответствующих переменных \mathbf{Z} .

в) Ортогональная регрессия. Если ошибки факторов не зависят друг от друга и от ошибок в уравнениях (которые в этом случае интерпетируются как ошибки изучаемой переменной), их дисперсии одинаковы и равны дисперсии ошибки изучаемой переменной, а между истинными значениями переменных имеется линейная зависимость, то можно использовать ортогональную регрессию. Возвращаясь к обозначениям 3-го раздела,

$$\begin{aligned} \hat{\mathbf{X}}\boldsymbol{\alpha} &= \boldsymbol{\varepsilon} \text{ и} \\ (\mathbf{M} - \lambda \mathbf{I}_n)\mathbf{a} &= \mathbf{0}, \quad \mathbf{a}'\mathbf{a} = 1. \end{aligned}$$

В этом случае матрица ковариации ошибок переменных имеет вид $\sigma^2 \mathbf{I}_n$. Если матрица ковариации ошибок есть $\sigma^2 \boldsymbol{\Omega}$, то применяется регрессия в метрике $\boldsymbol{\Omega}^{-1}$:

$$(\mathbf{M} - \lambda \boldsymbol{\Omega})\mathbf{a} = \mathbf{0}, \quad \mathbf{a}'\boldsymbol{\Omega}\mathbf{a} = 1.$$

Для доказательства проводится преобразование в пространстве переменных с помощью матрицы \mathbf{C} , такой, что $\boldsymbol{\Omega} = \mathbf{C}^{-1} \mathbf{C}^{-1}$, после которого матрица ковариации ошибок переменных приобретает вид $\sigma^2 \mathbf{I}_n$, и становится возможным применить обычную ортогональную регрессию.

Ортогональная регрессия при принятых гипотезах приводит к состоятельным оценкам параметров.

6.2. Фиктивные переменные

С помощью фиктивных или псевдо- переменных, принимающих дискретные, обычно, целые значения, в регрессию включают качественные факторы.

Уточнение обозначений:

\mathbf{Z} - $\mathbf{N} \times \mathbf{n}$ -матрица наблюдений за “обычными” независимыми факторами;

$\boldsymbol{\alpha}$ - \mathbf{n} -вектор-столбец параметров регрессии при этих факторах;

$$\mathbf{Z}^0 = \mathbf{1}_N;$$

$$\boldsymbol{\beta}^0 = \boldsymbol{\beta}.$$

В этих обозначениях уравнение регрессии записывается следующим образом:

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\alpha} + \mathbf{Z}^0\boldsymbol{\beta}^0 + \boldsymbol{\varepsilon}.$$

Пусть имеется один качественный фактор, принимающий два значения (например: “мужчина” и “женщина”, если речь идет о модели некоторой характеристики отдельных людей, или “годы войны” и “годы мира” - в модели, построенной на временных рядах наблюдений, которые охватывают периоды войны и мира, и т.д.). Ставится вопрос о том, влияет ли этот фактор на значение свободного члена регрессии.

$\tilde{\mathbf{Z}}^F = \{\mathbf{z}_{ij}^F\}$ – $N \times 2$ -матрица наблюдений за качественным фактором (матрица фиктивных переменных): \mathbf{z}_{i1}^F равен единице, если фактор в i -м наблюдении принимает 1-е значение, и нулю в противном случае; \mathbf{z}_{i2}^F равен единице, если фактор в i -м наблюдении принимает 2-е значение, и нулю в противном случае.

$\tilde{\boldsymbol{\beta}} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$ – 2-х компонентный вектор-столбец параметров при фиктивных переменных.

Исходная форма регрессии с фиктивными переменными:

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\alpha} + \mathbf{Z}^0\boldsymbol{\beta}^0 + \tilde{\mathbf{Z}}^F\tilde{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}.$$

Поскольку сумма столбцов матрицы $\tilde{\mathbf{Z}}^F$ равна \mathbf{Z}^0 , оценка параметров непосредственно по этому уравнению невозможна.

Проводится преобразование фиктивных переменных одним из двух способов.

а) В исходной форме регрессии исключается один из столбцов матрицы фиктивных переменных, в данном случае - первый.

$\overline{\mathbf{Z}}^F$ - матрица фиктивных переменных без первого столбца;

$$\overline{\mathbf{C}} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}.$$

Тогда эквивалентная исходной запись уравнения имеет вид:

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\alpha} + [\mathbf{Z}^0, \overline{\mathbf{Z}}^F]\overline{\mathbf{C}} \begin{bmatrix} \beta^0 \\ \tilde{\boldsymbol{\beta}} \end{bmatrix} + \boldsymbol{\varepsilon},$$

и после умножения матрицы $\overline{\mathbf{C}}$ справа на вектор параметров получается запись уравнения регрессии в которой отсутствует линейная зависимость между факторами-регрессорами:

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\alpha} + \mathbf{Z}^0\overline{\boldsymbol{\beta}}^0 + \overline{\mathbf{Z}}^F\overline{\boldsymbol{\beta}} + \boldsymbol{\varepsilon},$$

где $\overline{\boldsymbol{\beta}}^0 = \boldsymbol{\beta}^0 + \boldsymbol{\beta}_1$, $\overline{\boldsymbol{\beta}} = \boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$.

После оценки этих параметров можно определить значения исходных параметров $\boldsymbol{\beta}^0$ и $\tilde{\boldsymbol{\beta}}$, предполагая, что сумма параметров при фиктивных переменных (в данном случае $\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2$) равна нулю, т.е. влияние качественного фактора приводит к колебаниям вокруг общего уровня свободного члена:

$$\boldsymbol{\beta}_2 = \overline{\boldsymbol{\beta}}/2, \boldsymbol{\beta}_1 = -\boldsymbol{\beta}_2, \boldsymbol{\beta}^0 = \overline{\boldsymbol{\beta}}^0 + \boldsymbol{\beta}_2.$$

б) Предполагая, что сумма параметров при фиктивных переменных равна нулю, в исходной форме регрессии исключается один из этих параметров, в данном случае - первый.

β - вектор-столбец параметров при фиктивных переменных без первого элемента;

$$C = \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

Эквивалентная исходной запись уравнения принимает форму:

$$X = Z\alpha + Z^0\beta^0 + \tilde{Z}^F C\beta + \varepsilon,$$

и после умножения матрицы C слева на матрицу наблюдений за фиктивными переменными получается запись уравнения регрессии, в которой также отсутствует линейная зависимость между регрессорами:

$$X = Z\alpha + Z^0\beta^0 + Z^F\beta + \varepsilon.$$

После оценки параметров этого уравнения недостающая оценка параметра β_1 определяется из условия $\beta_1 = -\beta_2$.

Качественный фактор может принимать больше двух значений. Так, в классической модели выделения сезонных колебаний он принимает 4 значения в случае поквартальных наблюдений и 12 значений, если наблюдения проводились по месяцам. Матрица \tilde{Z}^F в этой модели имеет размерность, соответственно, $N \times 4$ или $N \times 12$.

Пусть в общем случае качественный фактор принимает k значений. Тогда:

матрица \tilde{Z}^F имеет размерность $N \times k$, вектор-столбец $\tilde{\beta}$ - размерность k , матрицы \bar{Z}^F и Z^F - $N \times (k-1)$, вектора-столбцы $\bar{\beta}$ и β - $k-1$;

$$k \times (k+1) \quad \text{матрица} \quad \bar{C} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & -1_{k-1} & I_{k-1} \end{bmatrix}, \quad k \times (k-1) \quad \text{матрица} \\ C = \begin{bmatrix} -1'_{k-1} \\ I_{k-1} \end{bmatrix}; \quad 1'_k \tilde{\beta} = 0, \quad \bar{C} \begin{bmatrix} \beta^0 \\ \tilde{\beta} \end{bmatrix} = \begin{bmatrix} \bar{\beta}^0 \\ \bar{\beta} \end{bmatrix}, \quad \tilde{Z}^F C = Z^F.$$

Можно показать, что

$$\begin{bmatrix} 1 & -1'_{k-1} \\ 0 & I_{k-1} - 1^{k-1} \end{bmatrix} \begin{bmatrix} \beta^0 \\ \beta \end{bmatrix} = \begin{bmatrix} \bar{\beta}^0 \\ \bar{\beta} \end{bmatrix}, \quad \text{или} \quad \begin{bmatrix} 1 & 1'_{k-1} (I_{k-1} - \frac{1}{k} 1^{k-1}) \\ 0 & I_{k-1} - \frac{1}{k} 1^{k-1} \end{bmatrix} \begin{bmatrix} \bar{\beta}^0 \\ \bar{\beta} \end{bmatrix} = \begin{bmatrix} \beta^0 \\ \beta \end{bmatrix},$$

где $1^{k-1} = 1_{k-1} 1'_{k-1}$ - $(k-1) \times (k-1)$ -матрица, состоящая из единиц; и далее показать, что результаты оценки параметров уравнения с фиктивными переменными при использовании обоих указанных подходов к устранению линейной зависимости факторов-регрессоров одинаковы.

После оценки регрессии можно применить t -критерий для проверки значимости влияния качественного фактора на свободный член уравнения.

Если k слишком велико и приближается к N , то на параметры при фиктивных переменных накладываются более жесткие ограничения (чем равенство нулю их суммы). Так, например, если наблюдения проведены в последовательные моменты

времени, и вводится качественный фактор “время”, принимающий особое значение в каждый момент времени, то $\tilde{Z}^F = I_N$, и обычно предполагается, что значение параметра в каждый момент времени (при фиктивной переменной каждого момента времени) больше, чем в предыдущий момент времени на одну и ту же величину. Тогда роль матрицы C играет N -вектор-столбец T , состоящий из чисел натурального ряда, начиная с 1, и $\tilde{\beta} = T\beta_T$, где β_T - скаляр. Уравнение регрессии с фактором времени имеет вид (эквивалентная исходной форма уравнения при использовании способа “б” исключения линейной зависимости фиктивных переменных):

$$X = Z\alpha + Z^0\beta^0 + T\beta_T + \varepsilon.$$

Метод фиктивных переменных можно использовать для проверки влияния качественного фактора на коэффициент регрессии при любом обычном факторе. Исходная форма уравнения, в которое вводится качественный фактор для параметра α_j , имеет следующий вид:

$$X = Z\alpha + Z^0\beta^0 + Z_j \otimes \tilde{Z}^F \tilde{\alpha}^j + \varepsilon,$$

где Z_j – j -й столбец матрицы Z ,

$\tilde{\alpha}^j$ - k -вектор-столбец параметров влияния качественного фактора на α_j ;

в векторе α j -я компонента теперь обозначается α_j^0 - средний уровень параметра α_j ;

\otimes - операция прямого произведения столбцов матриц.

Замечание

Прямое произведение матриц $A \otimes B$, имеющих размерность, соответственно, $m_A \times n_A$ и $m_B \times n_B$ есть матрица размерности $(m_A m_B) \times (n_A n_B)$ следующей структуры:

$$\begin{bmatrix} a_{11} B & . & . & a_{1n_A} B \\ . & . & . & . \\ . & . & . & . \\ a_{m_A 1} B & . & . & a_{m_A n_A} B \end{bmatrix}$$

Прямое произведение матриц обладает следующими свойствами:

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD), \text{ если произведения } AC \text{ и } BD \text{ имеют смысл,}$$

$$(A \otimes B)' = A' \otimes B', (A \otimes B)^{-1} = A^{-1} \otimes B^{-1}.$$

Прямое произведение столбцов матриц применимо к матрицам, имеющим одинаковое число строк, и осуществляется путем проведения операции прямого произведения последовательно с векторами-строками матриц.

Приоритет прямого произведения матриц выше, чем обычного матричного произведения.

При использовании способа “а” эквивалентная исходной форма уравнения имеет вид (форма “а”):

$$X = Z_{-j} \alpha_{-j} + Z^0 \beta^0 + Z_j \otimes [Z^0, \overline{Z}^F] \overline{C} \begin{bmatrix} \alpha_j^0 \\ \tilde{\alpha}^j \end{bmatrix} + \varepsilon,$$

где Z_{-j} - матрица Z без j -го столбца,

α_{-j} - вектр α без j -го элемента;

а в случае применения способа “б” (форма “б”):

$$X = Z\alpha + Z^0\beta^0 + Z_j \otimes \tilde{Z}^F C\alpha^j + \varepsilon.$$

Все приведенные выше структуры матриц и соотношения между матрицами и векторами сохраняются.

В уравнение регрессии можно включать более одного качественного фактора. В случае двух факторов, принимающих, соответственно, k_1 и k_2 значения, форма “б” уравнения записывается следующим образом:

$$X = Z\alpha + Z^0\beta^0 + Z^1\beta^1 + Z^2\beta^2 + \varepsilon,$$

где вместо “F,” в качестве индекса качественного фактора используется его номер.

Это уравнение может включать фиктивные переменные совместного влияния качественных факторов (взаимодействия факторов). В исходной форме компонента совместного влияния записывается следующим образом:

$$\tilde{Z}^1 \otimes \tilde{Z}^2 \tilde{\beta}^{12},$$

где $\tilde{\beta}^{12}$ - $k_1 \times k_2$ -вектор-столбец $(\beta_{11}^{12}, \dots, \beta_{1k_2}^{12}, \beta_{21}^{12}, \dots, \beta_{2k_2}^{12}, \dots, \beta_{k_1 1}^{12}, \dots, \beta_{k_1 k_2}^{12})'$,

а $\beta_{i_1 i_2}^{12}$ - параметр при фиктивной переменной, которая равна 1, если 1-й фактор принимает i_1 -е значение, а 2-й фактор - i_2 -е значение, и равна 0 в остальных случаях (вектором-столбцом наблюдений за этой переменной является $(k_1(i_1-1)+i_2)$ -й столбец матрицы $\tilde{Z}^1 \otimes \tilde{Z}^2$).

Как и прежде, вектор параметров, из которого исключены все компоненты, линейно выражаемые через остальные, обозначается β^{12} . Он имеет размерность $(k_1-1) \times (k_2-1)$ и связан с исходным вектором параметров таким образом:

$$\tilde{\beta}^{12} = C^1 \otimes C^2 \beta^{12},$$

где C^1 и C^2 - матрицы размерности $k_1 \times (k_1-1)$ и $k_2 \times (k_2-1)$, имеющие описанную выше структуру (матрица C).

Теперь компоненту совместного влияния можно записать следующим образом:

$$(\tilde{Z}^1 \otimes \tilde{Z}^2)(C^1 \otimes C^2)\beta^{12} = (\tilde{Z}^1 C^1) \otimes (\tilde{Z}^2 C^2)\beta^{12} = Z^1 \otimes Z^2 \beta^{12} = Z^{12} \beta^{12},$$

а уравнение, включающее эту компоненту (форма “б”) -

$$X = Z\alpha + Z^0\beta^0 + Z^1\beta^1 + Z^2\beta^2 + Z^{12}\beta^{12} + \varepsilon.$$

В общем случае имеется L качественных факторов, j-й фактор принимает k_j значений. Пусть упорядоченное множество $\{1, 2, \dots, L\}$ обозначается F, а J - его подмножества. Общее их количество, включая пустое подмножество, равно 2^L . Каждому такому подмножеству взаимно однозначно соответствует число, например, в системе исчисления с основанием $\max_j k_j$, и их можно упорядочить по

возрастанию этих чисел. Если пустое подмножество обозначить 0, то можно записать $J = 0, 1, \dots, L, \{1, 2\}, \dots, \{1, L\}, \{2, 3\}, \dots, \{1, 2, 3\}, \dots, F$. Тогда уравнение регрессии записывается следующим образом:

$$X = Z\alpha + \sum_{J=0}^F \tilde{Z}^J \tilde{\beta}^J + \varepsilon = Z\alpha + \sum_{J=0}^F \tilde{Z}^J C^J \beta^J + \varepsilon = Z\alpha + \sum_{J=0}^F Z^J \beta^J + \varepsilon,$$

где $\tilde{Z}^J = \prod_{j \in J} \tilde{Z}^j$, $C^J = \prod_{j \in J} C^j$ при $j > 0$; $C^0 = 1$. Выражение $j \in J$

означает, что j принимает значения последовательно с 1-го по последний элемент подмножества J .

Очевидно, что приведенная выше запись уравнения для $L = 2$ является частным случаем данной записи.

Если $p(J)$ - количество элементов в подмножестве J , то

$\tilde{Z}^J \tilde{\beta}^J$ или $Z^J \beta^J$ - J -е эффекты, эффекты $p(J)$ -го порядка, при $p(J) = 1$ - главные эффекты, при $p(J) > 1$ - эффекты взаимодействия, эффекты совместного влияния или совместные эффекты.

$\tilde{\beta}^J$ или β^J - параметры соответствующих J -х эффектов или также сами эти эффекты.

6.3. Дисперсионный анализ

Рассматривается частный случай уравнения регрессии с фиктивными переменными, когда оно включает только такие (фиктивные) переменные, и для каждого сочетания значений факторов имеется одно и только одно наблюдение за изучаемой переменной. Тогда $N = \prod_{j \in F} k_j$ и уравнение имеет вид:

$$X = \sum_{J=0}^F Z^J \beta^J = Z\beta,$$

в котором отсутствует вектор ошибок ε , т.к. при учете эффектов всех порядков их сумма в точности равняется X .

Матрица Z имеет размерность $N \times N$ и она не вырождена. Поэтому $\beta = Z^{-1}X$. Но чтобы получить общие результаты, имеющие значение и для частных моделей, в которых эффекты высоких порядков принимаются за случайную ошибку, ниже используется техника регрессионного анализа.

Это - регрессионная модель полного (учитываются эффекты всех порядков) одномерного (изучаемая переменная единственна) многофакторного дисперсионного анализа без повторений (для каждого сочетания значений факторов есть одно наблюдение).

Обычному линейному индексу $i = \overline{1, N}$ компонент вектора X можно поставить в соответствие мультииндекс I , принимающий значения из множества $\prod_{j \in F} \{\overline{1, k_j}\}$, так что, если $I = \{i_1, i_2, \dots, i_L\}$, то

$i = (\dots ((i_1 - 1)k_2 + (i_2 - 1))k_3 + \dots)k_L + i_L$, и - при этом - обозначения x_i и

x_I эквивалентны. При таком соответствии обычного индекса и мультииндекса в линейной последовательности значений мультииндекса быстрее меняются его младшие компоненты (с большим порядковым номером).

$N^J = \prod_{j \in J} k_j$, если $j > 0$, и $N^0 = 1$ - количество столбцов в матрице \tilde{Z}^J ;

$N_-^J = \prod_{j \in J} (k_j - 1)$, если $j > 0$, и $N_-^0 = 1$ - количество столбцов в матрице

Z^J ; очевидно, что $N^F = \sum_{J=0}^F N_-^J = N$;

$I^J = \{i_1, \dots, i_{p(J)}\}$ - мультииндекс с множеством значений $\prod_{j \in J} \overline{\{1, k_j\}}$;

$I = I^F$.

$Mb = m$ - система нормальных уравнений,

где M - $N \times N$ -матрица, b и m - N -вектора-столбцы и, как обычно,

$$M = \frac{1}{N} Z' Z, \quad m = \frac{1}{N} Z' X.$$

При выбранном порядке следования значений факторов от наблюдения к наблюдению (быстрее меняют свои значения более младшие факторы)

$\tilde{Z}^J = \prod_{j \in F} \otimes \xi_j$ где ξ_j есть I_{k_j} , если $j \in J$, или 1_{k_j} , в противном случае.

Тогда

$Z^J = \prod_{j \in F} \otimes \xi_j$ где ξ_j есть C^j , если $j \in J$, или 1_{k_j} , в противном случае, и

далее

$Z^{\bar{J}'} Z^J = 0$, если $\bar{J} \neq J$, т.е. переменные разных эффектов ортогональны друг другу,

$$M^J = \frac{1}{N} Z^{J'} Z^J = \frac{1}{N^J} C^{J'} C^J = \prod_{j \in J} \otimes M^j, \quad M^0 = 1;$$

$$m^J = \frac{1}{N} Z^{J'} X = \frac{1}{N} C^{J'} \tilde{Z}^{J'} X = \frac{1}{N^J} C^{J'} X^J,$$

где $X^J = \frac{N^J}{N} \tilde{Z}^{J'} X$ - N^J -вектор-столбец средних по сочетаниям значений

факторов J с мультииндексом компонент I^J ($x_{I^J}^J$ является средним значением x по тем наблюдениям, в которых 1-й фактор из множества J принимает i_1 -е значение, 2-й - i_2 -е значение и т.д.); $X^0 = \bar{x}$, $X^F = X$.

M - блочно-диагональная матрица $\{M^J\}$, m - вектор-столбец $\{m^J\}$.

После решения системы нормальных уравнений и перехода к “полным” векторам параметров эффектов получается следующее:

$$\tilde{b}^J = C^J (C^{J'} C^J)^{-1} C^{J'} X^J = B^J X^J = \left(\prod_{j \in J} \otimes B^j \right) X^J,$$

где $B^j = I_{k_j} - \frac{1}{k_j} 1^{k_j}$ (как и прежде, $1^{k_j} = 1_{k_j} 1_{k_j}'$), $B^0 = 1$.

Параметры разных эффектов $\tilde{\mathbf{b}}^{\mathbf{J}}$ (разных по \mathbf{J}) не зависят друг от друга, и исключение из уравнения некоторых из них не повлияет на значения параметров оставшихся эффектов.

Чтобы получить более “прозрачные” формулы для определения параметров эффектов, следует ввести понятие сопоставимых векторов этих параметров.

Если $\bar{\mathbf{J}} \subset \mathbf{J}$, то $\tilde{\mathbf{b}}^{\bar{\mathbf{J}}\mathbf{J}} = \mathbf{B}^{\bar{\mathbf{J}}\mathbf{J}} \mathbf{X}^{\mathbf{J}}$ - $\mathbf{N}^{\mathbf{J}}$ -вектор-столбец параметров $\bar{\mathbf{J}}$ -го эффекта, сопоставимый с вектором $\tilde{\mathbf{b}}^{\mathbf{J}}$: он имеет ту же размерность, что и $\tilde{\mathbf{b}}^{\mathbf{J}}$, и каждая компонента вектора $\tilde{\mathbf{b}}^{\bar{\mathbf{J}}}$ повторена в нем $\frac{\mathbf{N}^{\mathbf{J}}}{\mathbf{N}^{\bar{\mathbf{J}}}}$ раз - так, что любой компоненте $\mathbf{b}_{\mathbf{I}^{\mathbf{J}}}^{\mathbf{J}}$ вектора $\tilde{\mathbf{b}}^{\mathbf{J}}$ в векторе $\tilde{\mathbf{b}}^{\bar{\mathbf{J}}\mathbf{J}}$ соответствует компонента $\mathbf{b}_{\mathbf{I}^{\bar{\mathbf{J}}}}^{\bar{\mathbf{J}}}$, для которой $\mathbf{I}^{\bar{\mathbf{J}}}$ является подмножеством тех же элементов $\mathbf{I}^{\mathbf{J}}$, что и $\bar{\mathbf{J}}$ по отношению к \mathbf{J} .

В этом выражении для сопоставимых векторов параметров эффектов $\mathbf{B}^{\bar{\mathbf{J}}\mathbf{J}} = \prod_{j \in \bar{\mathbf{J}}} \otimes \xi_j$, где ξ_j равен $\mathbf{B}^{\mathbf{j}}$, если $j \in \bar{\mathbf{J}}$, или $\frac{1}{\mathbf{k}_j} \mathbf{1}^{\mathbf{k}_j}$, в противном случае ($\mathbf{B}^{0\mathbf{J}} = \frac{1}{\mathbf{N}^{\mathbf{J}}} \mathbf{1}^{\mathbf{N}^{\mathbf{J}}}$, $\mathbf{B}^{\mathbf{J}\mathbf{J}} = \mathbf{B}^{\mathbf{J}}$).

Эти матрицы обладают следующим свойством: $\sum_{\bar{\mathbf{J}}=0}^{\mathbf{J}} \mathbf{B}^{\bar{\mathbf{J}}\mathbf{J}} = \mathbf{I}_{\mathbf{N}^{\mathbf{J}}}$, откуда получается выражение

$$\mathbf{X}^{\mathbf{J}} = \sum_{\bar{\mathbf{J}}=0}^{\mathbf{J}} \tilde{\mathbf{b}}^{\bar{\mathbf{J}}\mathbf{J}} = \sum_{\substack{\bar{\mathbf{J}} \neq \mathbf{J} \\ \bar{\mathbf{J}} \subset \mathbf{J}}} \tilde{\mathbf{b}}^{\bar{\mathbf{J}}\mathbf{J}} + \tilde{\mathbf{b}}^{\mathbf{J}}$$

для рекуррентного расчета параметров эффектов (например, если известны \mathbf{b}^0 , $\mathbf{b}_{i_1}^1$, $\mathbf{b}_{i_1 i_2}^2$, то $\mathbf{b}_{i_1 i_2}^{12} = \mathbf{x}_{i_1 i_2}^{12} - \mathbf{b}^0 - \mathbf{b}_{i_1}^1 - \mathbf{b}_{i_2}^2$).

При $\mathbf{J} = \mathbf{F}$ это выражение представляет собой другую форму записи основного уравнения регрессии:

$$\mathbf{X} = \sum_{\mathbf{J}=0}^{\mathbf{F}} \tilde{\mathbf{b}}^{\mathbf{J}\mathbf{F}}, \text{ т.е. } \mathbf{Z}^{\mathbf{J}} \mathbf{b}^{\mathbf{J}} = \tilde{\mathbf{b}}^{\mathbf{J}\mathbf{F}}.$$

$\mathbf{s}_{\mathbf{x}}^2 = \sum_{\mathbf{J}=1}^{\mathbf{F}} \mathbf{s}_{\mathbf{J}}^2$ - основное тождество дисперсионного анализа, показывающее распределение общей дисперсии изучаемой величины по факторам и их взаимодействиям,

где $\mathbf{s}_{\mathbf{J}}^2 = \frac{1}{\mathbf{N}^{\mathbf{J}}} \mathbf{X}^{\mathbf{J}/\sim \mathbf{J}} \tilde{\mathbf{b}}^{\mathbf{J}}$ - дисперсия, объясненная совместным влиянием факторов \mathbf{J} ; представляет собой сумму квадратов с $\mathbf{N}_{-}^{\mathbf{J}}$ степенями свободы.

Все эти дисперсии не зависят друг от друга. Если совместное влияние факторов $\bar{\mathbf{J}}$ так же существенно (или не существенно) как и факторов \mathbf{J} , то статистика

$$\frac{s_{\bar{J}}^2/N_{\bar{J}}}{s_J^2/N_J} \quad (\text{предполагается, что она больше единицы})$$

имеет $F_{N_{\bar{J}}, N_J}$ -распределение (предполагается, что \mathbf{X} нормально распределено).

Этот факт можно использовать для проверки гипотез о сравнительной существенности факторов и их взаимодействий.

Обычно эффекты высоких порядков отождествляют со случайной ошибкой. Уравнение регрессии приобретает свою обычную форму и можно воспользоваться t - и F -критериями для проверки значимости отдельных факторов и их взаимодействий. Важно, что оценки оставшихся в уравнении эффектов при этом не меняются.

Переходя к более общему и более сложному случаю модели дисперсионного анализа с повторениями, полезно вспомнить следующее. Если в модели регрессионного анализа

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$$

несколько строк матрицы \mathbf{Z} одинаковы, то можно перейти к сокращенной модели, в которой из всех этих строк оставлена одна, а в качестве соответствующей компоненты вектора \mathbf{X} взято среднее по этим наблюдениям с одинаковыми значениями независимых факторов. Т.е. совокупность наблюдений с одинаковыми значениями независимых факторов заменяется одним групповым наблюдением. При исходной гипотезе $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\mathbf{I}$ дисперсия остатка по этому наблюдению равна $n^g\sigma^2$, где n^g - количество замененных наблюдений, и значения переменных в групповом наблюдении должны быть умножены на $\sqrt{n^g}$ (в соответствии с ОМНК). Значения оценок параметров по исходной и сокращенной модели будут одинаковыми, но полная $(\hat{\mathbf{X}}'\hat{\mathbf{X}})$ и остаточная $(\mathbf{e}'\mathbf{e})$ суммы квадратов в исходной модели будут больше, чем в сокращенной на сумму квадратов отклонений переменных \mathbf{X} по исключенным наблюдениям от своей средней.

Пусть теперь рассматривается регрессионная модель одномерного однофакторного дисперсионного анализа с повторениями:

$$\mathbf{X} = [\mathbf{Z}^0, \tilde{\mathbf{Z}}] \begin{bmatrix} \beta^0 \\ \tilde{\beta} \end{bmatrix} + \boldsymbol{\varepsilon}.$$

Фактор принимает k значений, и для каждого i -го значения существует n_i наблюдений (n_i повторений), т.е. исходная совокупность \mathbf{X} разбита по какому-то признаку на k групп, причем сначала в ней идут наблюдения по 1-й группе, потом - по 2-й и т.д..

$$\mathbf{N} = \sum_{i=1}^k n_i; \quad \tilde{\mathbf{Z}} - \mathbf{N} \times k\text{-матрица структуры} \begin{bmatrix} 1_{n_1} & 0 & \cdot & 0 \\ 0 & 1_{n_2} & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & 1_{n_k} \end{bmatrix}.$$

Всем повторениям в матрице \mathbf{Z} соответствуют одинаковые строки, поэтому можно перейти к сокращенной модели.

\bar{x}_i - среднее и s_i^2 - дисперсия по i -й группе; $s_e^2 = \frac{1}{N} \sum_{i=1}^k n_i s_i^2$ - суммарная дисперсия по группам. Сокращенная модель имеет следующий вид:

$$\sqrt{n_i} \bar{x}_i = \sqrt{n_i} (\beta^0 + \beta_i), \quad i = \overline{1, k}.$$

При естественном требовании $b^0 = \bar{x}$, которое эквивалентно $\sum_{i=1}^k n_i b_i = 0$,

матрица C имеет вид

$$\begin{bmatrix} -\frac{n_2}{n_1} & -\frac{n_3}{n_1} & \cdot & \cdot & -\frac{n_k}{n_1} \\ 1 & 0 & \cdot & \cdot & 0 \\ 0 & 1 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & 1 \end{bmatrix} \text{ и } b_i = \bar{x}_i - \bar{x}.$$

$s_q^2 = \frac{1}{N} \sum_{i=1}^k n_i b_i^2$ - объясненная дисперсия, равная полной дисперсии в сокращенной модели.

Полная дисперсия в исходной модели распадается на две части:

$$s_x^2 = s_q^2 + s_e^2$$

- объясненную и остаточную, или в терминах дисперсионного анализа - межгрупповую и внутригрупповую дисперсии, которые имеют, соответственно, k и $N-k-1$ степеней свободы. Применяя F -критерий, можно оценить статистическую значимость использования данной группировки в целом или выделения отдельных групп.

Теперь рассматривается общий случай L -факторной модели.

В этом случае N больше N^F на общее число повторений по всем сочетаниям значений факторов. Пусть

n_I - число наблюдений при I -м сочетании значений факторов;
 $n_I \geq 1, \sum_I n_I = N$;

x_I - среднее значение и s_I^2 - дисперсия наблюдений при I -м сочетании;

$s_e^2 = \frac{1}{N} \sum_I n_I s_I^2$ - суммарная внутригрупповая или остаточная дисперсия для

исходной модели с $N-N^F-1$ степенями свободы.

Сокращенная модель имеет вид:

$$n^{0.5} X = n^{0.5} Z \beta,$$

где n - диагональная N^F -матрица $\{n_I\}$;

X - N^F -вектор-столбец $\{x_I\}$;

Z, β - аналогичны L -факторной модели без повторений.

Пусть далее

$$\tilde{M} = \frac{1}{N} n,$$

$N^{\bar{J}} \times N^J$ -матрица $\tilde{M}^{\bar{J}J} = \tilde{Z}^{\bar{J}} / \tilde{M} \tilde{Z}^J$, в частности $N \tilde{M}^{JJ}$ - диагональная N^J - матрица $\{n_{1j}^J\}$, где n_{1j}^J - количество наблюдений при I^J -м сочетании значений факторов J ($\tilde{M}^{FF} = \tilde{M}$);

$N^{\bar{J}} \times N^J$ -матрица $M^{\bar{J}J} = C^{\bar{J}} / \tilde{M}^{\bar{J}J} C^J$,

N^J -вектор-столбец $m^J = C^J / \tilde{M}^{JJ} X^J$,

где $X^J = \tilde{M}^{JJ-1} \tilde{Z}^J / \tilde{M} X$ - N^J -вектор-столбец средневзвешенных x по сочетаниям значений факторов J .

Матрица M и вектор m системы нормальных уравнений для b составляются естественным образом из блоков $M^{\bar{J}J}$ и m^J .

Формулы для M^J (в данном случае M^{JJ}), m^J и X^J , приведенные для модели без повторений, являются частным случаем этих формул при $n = I_{N^F}$.

$s_q^2 = m' M^{-1} m - \bar{x}^2 = X' \tilde{M} (\tilde{M}^{-1} - 1^{N^F}) \tilde{M} X$ - полная дисперсия в сокращенной модели или объясненная дисперсия в исходной модели.

Разные эффекты могут оставаться ортогональными ($M^{\bar{J}J} = 0$ при $\bar{J} \neq J$) в одном специальном случае, когда каждый более младший фактор делит все выделенные до него подгруппы в одинаковых пропорциях, т.е. $\tilde{M} = \prod_{j \in F} \tilde{M}^{jj}$ (в

частности, когда количество повторений n_i для всех сочетаний I одинаково). В этом случае для ортогональности эффектов достаточно матрицы C^j выбрать так, чтобы $1'_{k_j} \tilde{M}^j C^j = 0$. Эти требования удовлетворяются, если данные матрицы обладают описанной выше (для однофакторной модели с повторениями) структурой:

$$C^j = \begin{bmatrix} -c^j \\ I_{k_j-1} \end{bmatrix}, \text{ где } c^j = \frac{1}{n_1^j} (n_2^j, \dots, n_{k_j}^j).$$

Такие матрицы обобщают структуру матриц C^j модели без повтрений.

Для этого специального случая можно построить формулы решения задачи дисперсионного анализа, обобщающие приведенные выше формулы для модели без повторений.

В общем случае указанный выбор матриц C^j обеспечивает равенство нулю только M^{0j} . Особым выбором C^J ($p(J) > 1$) можно добиться равенства нулю еще некоторых блоков общей матрицы M .

Матрица C^J не обязательно должна равняться прямому произведению C^j по $j \in J$. Она должна быть размерности $N^J \times N^J$ и иметь ранг N^J , т.е., например,

обладать структурой $\begin{bmatrix} -c^J \\ I_{N^J} \end{bmatrix}$, где c^J - $(N^J - N^J_-) \times N^J_-$ -матрица. Поэтому для

определения этой матрицы необходимо иметь $(N^J - N_-^J) \times N_-^J$ условий. Поскольку

$$N^J - N_-^J = \sum_{\substack{\bar{J} \subset J \\ \bar{J} \neq J}} N_-^{\bar{J}},$$

нужное количество условий содержат требования

$$M^{\bar{J}J} = \frac{1}{N} C^{\bar{J}} / \tilde{M}^{\bar{J}J} C^J = 0$$

для всех $\bar{J} \subset J, \bar{J} \neq J$, включая пустое множество $\bar{J} = 0$ ($C^0 = 1$).

Таким образом, матрицы C^J всегда можно определить так, чтобы эффекты нулевого и высшего порядков были ортогональны друг с другом и с остальными эффектами, и, в частности, $b^0 = \bar{x}$.

Дисперсия s_q^2 в общем случае не делится на факторные дисперсии, как это было в модели без повторений; точно в ней выделяется только дисперсия эффектов высшего порядка (при указанном выборе C^J):

$$s_F^2 = X' \tilde{M} C^F (C^F / \tilde{M} C^F)^{-1} C^F / \tilde{M} X,$$

и для нее непосредственно можно проверить нулевую гипотезу с помощью F -критерия

$$\frac{s_F^2 / N_-^F}{s_e^2 / (N - N^F - 1)}.$$

Нулевые гипотезы для остальных факторных дисперсий имеют вид $\beta^J = 0$, и в числителе F -статистики помещается величина

$$b^J / (M^{JJ^{-1}})^{-1} b^J / N_-^J,$$

где $M^{JJ^{-1}}$ - соответствующий блок матрицы M^{-1} ,
а в знаменателе -

$$s_e^2 / (N - N^F - 1) \quad \text{или} \quad (s_e^2 + s_F^2) / (N + N_-^F - N^F - 1) \quad - \quad \text{если}$$

нулевая гипотеза для s_F^2 не отвергается.

Теоретические вопросы и задания

1(*). Доказать смещенность МНК-оценок в случае наличия ошибок в независимых переменных.

2. Почему, если известна оценка W ковариационной матрицы ошибок независимых переменных, то приведенная формула расчета оценок параметров простой регрессии обеспечивает их несмещенность?

3. Вывести формулу оценки Вальда углового коэффициента регрессии.

4(*). Почему при наличии ошибок во всех переменных применима ортогональная регрессия? Каким образом в этом случае регрессия в метрике Ω^{-1} играет роль взвешенной регрессии?

5. Для модели с фиктивными переменными вывести формулы, связывающие параметры $\tilde{\beta}$, $\bar{\beta}$ и β в общем случае.

6(*). Показать эквивалентность обоих приведенных способов устранения линейной зависимости между фиктивными переменными в исходной форме уравнения регрессии.

7. Оценка параметров систем уравнений

7.1. Невзаимозависимые системы

$\hat{\mathbf{x}}, \boldsymbol{\varepsilon}$ - \mathbf{k} -вектора-строки центрированных значений изучаемых (эндогенных) переменных и их случайных ошибок; $\mathbf{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\mathbf{E}(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{\Omega}$;

$\hat{\mathbf{z}}$ - \mathbf{n} -вектор-строка центрированных значений независимых факторов (экзогенных переменных);

\mathbf{A} - $\mathbf{n} \times \mathbf{k}$ -матрица коэффициентов регрессии;

$\hat{\mathbf{x}} = \hat{\mathbf{z}}\mathbf{A} + \boldsymbol{\varepsilon}$ - система уравнений регрессии;

$\hat{\mathbf{X}} = \hat{\mathbf{Z}}\mathbf{A} + \boldsymbol{\varepsilon}$ - та же система по \mathbf{N} наблюдениям; в каждом наблюдении матожидание ошибок равно нулю, их матрица ковариации одинакова (равна $\sigma^2\boldsymbol{\Omega}$) и они не скоррелированы по наблюдениям.

$$\mathbf{A} = \mathbf{M}_{\mathbf{ZZ}}^{-1} \mathbf{M}_{\mathbf{ZX}},$$

где $\mathbf{M}_{\mathbf{ZZ}} = \frac{1}{\mathbf{N}} \hat{\mathbf{Z}}' \hat{\mathbf{Z}}$, $\mathbf{M}_{\mathbf{ZX}} = \frac{1}{\mathbf{N}} \hat{\mathbf{Z}}' \hat{\mathbf{X}}$, т.е. факт скоррелированности

ошибок разных изучаемых переменных ($\boldsymbol{\Omega} \neq \mathbf{I}_k$) не создает дополнительных проблем, и уравнения системы могут оцениваться по отдельности с помощью обычного МНК.

Пусть для коэффициентов матрицы \mathbf{A} имеются априорные ограничения, и эта матрица имеет, например, следующую структуру:

$$\begin{bmatrix} \mathbf{a}_1 & \mathbf{0} & . & . & \mathbf{0} \\ \mathbf{0} & \mathbf{a}_2 & . & . & \mathbf{0} \\ . & . & . & . & . \\ . & . & . & . & . \\ \mathbf{0} & \mathbf{0} & . & . & \mathbf{a}_k \end{bmatrix},$$

где \mathbf{a}_i - \mathbf{n}_i -вектор-столбец коэффициентов в i -м уравнении (для i -й изучаемой переменной); $\sum_{i=1}^k \mathbf{n}_i = \mathbf{n}$. Т.е. для каждой изучаемой переменной имеется свой набор

объясняющих факторов с $\mathbf{N} \times \mathbf{n}_i$ -матрицей наблюдений $\hat{\mathbf{Z}}_i$ ($\hat{\mathbf{Z}} = [\hat{\mathbf{Z}}_1, \hat{\mathbf{Z}}_2, \dots, \hat{\mathbf{Z}}_k]$), и система уравнений записывается как совокупность внешне не связанных между собой уравнений:

$$\hat{\mathbf{X}}_i = \hat{\mathbf{Z}}_i \mathbf{a}_i + \boldsymbol{\varepsilon}_i, \quad i = \overline{1, k}.$$

Поскольку ошибки скоррелированы, правильная оценка параметров регрессии дается решением следующих уравнений:

$$\sum_{j=1}^k \omega_{ij}^{-1} \mathbf{M}_{ij} \mathbf{a}_j = \sum_{j=1}^k \omega_{ij}^{-1} \mathbf{m}_{ij}, \quad i = \overline{1, k},$$

где $M_{ij} = \frac{1}{N} \hat{Z}_i' \hat{Z}_j$, $m_{ij} = \frac{1}{N} \hat{Z}_i' \hat{X}_j$, ω_{ij}^{-1} - элемент матрицы Ω^{-1} .

Эта оценка совпадает с обычной МНК-оценкой $a_i = M_{ii}^{-1} m_{ii}$, если матрица Ω диагональна.

7.2. Взаимозависимые или одновременные уравнения. Проблема идентификации.

Уравнения регрессии записываются в форме без свободного члена.

X - $N \times k$ -матрица наблюдений за изучаемыми переменными x ;

Z - $N \times (n+1)$ -матрица наблюдений за независимыми факторами z ;

B - $k \times k$ -матрица параметров регрессии при изучаемых переменных; $|B| \neq 0$ и $\beta_{ii} = 1$ - условия нормализации, т.е. предполагается, что в конечном счете в левой части i -го уравнения остается только i -я переменная, а остальные изучаемые переменные переносятся в правую часть;

A - $(n+1) \times k$ -матрица параметров регрессии при независимых факторах;

ϵ - $N \times k$ -матрица значений случайных ошибок ϵ по наблюдениям;

$xB = zA + \epsilon$, или $XB = ZA + \epsilon$ - структурная форма системы уравнений регрессии;

$x = zAB^{-1} + \epsilon B^{-1}$, или $X = ZAB^{-1} + \epsilon B^{-1}$ - приведенная форма системы;

$D = AB^{-1}$ - $(n+1) \times k$ -матрица параметров регрессии приведенной формы.

Для их оценки используется МНК: $D = (Z'Z)^{-1} Z'X$.

$$DB - A = 0 \text{ или } WH = 0,$$

где $(n+1) \times (n+k+1)$ -матрица $W = [D, I_{n+1}]$,

$$(n+k+1) \times k\text{-матрица } H = \begin{bmatrix} B \\ -A \end{bmatrix},$$

- условия для оценки параметров структурной формы.

В общем случае этих условий недостаточно. Необходимы дополнительные условия. Пусть для параметров i -го уравнения имеется дополнительно r_i условий:

$$R_i h_i = 0,$$

где R_i - $r_i \times (n+k+1)$ -матрица дополнительных условий;

h_i - $(n+k+1)$ -вектор-столбец $\begin{bmatrix} B_i \\ -A_i \end{bmatrix}$ параметров i -го уравнения - i -й

столбец матрицы H .

$$\begin{bmatrix} W \\ R_i \end{bmatrix} h_i = W_i h_i = 0 \text{ - общие условия для определения структурных}$$

параметров i -го уравнения, где W_i - $(n+r_i+1) \times (n+k+1)$ -матрица.

Они позволяют определить искомые параметры с точностью до постоянного множителя (с точностью до выполнения условий нормализации $\beta_{II} = 1$), если ранг матрицы W_1 равен $n+k$. Для этого необходимо, чтобы $r_1 \geq k-1$; необходимо и достаточно, чтобы ранг матрицы $R_1 H$ равнялся $k-1$.

l -е уравнение не идентифицировано, если $r_1 < k-1$; оно точно идентифицировано, если $r_1 = k-1$ и ранг W_1 равен $n+k$; сверхидентифицировано, если $r_1 > k-1$ и строки R_1 линейно не зависимы.

Обычно строки матрицы R_1 являются ортами, т.е. дополнительные ограничения исключают некоторые переменные из структурной формы. Тогда, если k_1 и n_1 - количества, соответственно, изучаемых переменных и независимых факторов в l -м уравнении, то для его идентификации необходимо, чтобы $k_1 + n_1 \leq n+1$.

Дальнейшее изложение ведется в предположении, что строки матрицы R_1 - орты.

7.3. Оценка параметров отдельного уравнения

X^1 - $N \times k_1$ -матрица наблюдений за изучаемыми переменными x^1 , входящими в l -е уравнение;

X_1 - N -вектор-столбец наблюдений за l -й переменной x_1 ;

X_-^1 - $N \times (k_1-1)$ -матрица X^1 без столбца X_1 наблюдений за x_-^1 ;

β^1 - k_1 -вектор-столбец параметров при изучаемых переменных в l -м уравнении;

β_1 - (k_1-1) -вектор-столбец β^1 с обратным знаком и без l -го элемента (без элемента $\beta_{II} = 1$);

Z^1 - $N \times (n_1+1)$ -матрица наблюдений за независимыми факторами z^1 , входящими в l -е уравнение;

α_1 - (n_1+1) -вектор-столбец параметров при этих факторах;

ϵ_1 - N -вектор-столбец остатков ϵ_1 в l -м уравнении по наблюдениям;

$X^1 \beta^1 = Z^1 \alpha_1 + \epsilon_1$ или $X_1 = X_-^1 \beta_1 + Z^1 \alpha_1 + \epsilon_1$ - l -е уравнение регрессии.

Применение обычного МНК к этому уравнению дает в общем случае смещенные оценки.

Если данное уравнение точно идентифицировано, то для оценки его параметров можно использовать косвенный метод (КМ) наименьших квадратов. С помощью МНК оцениваются параметры приведенной формы системы уравнений, через которые однозначно выражаются структурные параметры данного уравнения. Можно записать уравнения для этой оценки. Действительно, условия

$$R_1 H_1 = 0$$

эквивалентны

$$T_1^B B_1 = \beta^1, \quad T_1^A A_1 = \alpha_1,$$

где $\mathbf{T}_1^{\mathbf{B}}$ - $\mathbf{k}_1 \times \mathbf{k}$ -матрица, полученная из \mathbf{I}_k вычеркиванием нужных строк;

$\mathbf{T}_1^{\mathbf{A}}$ - аналогичная $(\mathbf{n}_1 + 1) \times (\mathbf{n} + 1)$ -матрица для \mathbf{A}_1 .

Тогда для \mathbf{B}_1 и \mathbf{A}_1 , удовлетворяющим требуемым условиям, выполняется следующее:

$$\mathbf{T}_1^{\mathbf{B}'} \boldsymbol{\beta}^1 = \mathbf{B}_1, \quad \mathbf{T}_1^{\mathbf{A}'} \boldsymbol{\alpha}_1 = \mathbf{A}_1,$$

и требования $\mathbf{W}\mathbf{H}_1 = \mathbf{0}$ можно записать в форме (переходя к обозначениям оценок соответствующих величин)

$$(\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{X}^1 \mathbf{b}^1 - \mathbf{T}_1^{\mathbf{A}'} \mathbf{a}_1 = \mathbf{0}, \quad (\text{т.к. } \mathbf{D} = (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{X} \text{ и } \mathbf{X}\mathbf{T}_1^{\mathbf{B}'} = \mathbf{X}^1)$$

$$\text{или} \quad \mathbf{d}_1 = \mathbf{D}_1 \mathbf{b}_1 + \mathbf{T}_1^{\mathbf{A}'} \mathbf{a}_1,$$

где $(\mathbf{n} + 1)$ -вектор-столбец $\mathbf{d}_1 = (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{X}_1$ (1-й столбец матрицы \mathbf{D});

$(\mathbf{n} + 1) \times (\mathbf{k}_1 - 1)$ -матрица $\mathbf{D}_1 = (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{X}_-^1$ (матрица, составленная из столбцов матрицы \mathbf{D} , соответствующих переменным \mathbf{x}_-^1).

Это - система уравнений для нахождения искомых параметров. Она имеет единственное решение в случае точной идентификации уравнения, т.е., если ее матрица

$$\begin{bmatrix} \mathbf{D}_1, \mathbf{T}_1^{\mathbf{A}'} \end{bmatrix}$$

квадратна, размерности $\mathbf{n} + 1$ и не вырождена (необходимое и достаточное условие точной идентификации уравнения).

Для сверхидентифицированного уравнения можно применить **двухшаговый метод** (2М) наименьших квадратов.

На 1-м шаге с помощью МНК оцениваются параметры приведенной формы для переменных \mathbf{X}_-^1 :

$$\mathbf{X}_-^1 = \mathbf{Z}\mathbf{D}_1 + \mathbf{V}^1,$$

где \mathbf{V}^1 - $\mathbf{N} \times (\mathbf{k}_1 - 1)$ -матрица остатков по уравнениям; и определяются расчетные значения этих переменных (“очищенные” от ошибок):

$$\mathbf{X}_-^{1^c} = \mathbf{Z}\mathbf{D}_1.$$

На 2-м шаге с помощью МНК оцениваются искомые параметры структурной формы из уравнения:

$$\mathbf{X}_1 = \mathbf{X}_-^{1^c} \mathbf{b}_1 + \mathbf{Z}^1 \mathbf{a}_1 + \mathbf{e}_1.$$

Можно определить единый оператор 2М-оценивания. Поскольку

$$\mathbf{X}_-^{1^c} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{X}_-^1 = \mathbf{F}\mathbf{X}_-^1 \text{ и } \mathbf{X}_1 = [\mathbf{F}\mathbf{X}_-^1, \mathbf{Z}^1] \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{a}_1 \end{bmatrix} + \mathbf{e}_1,$$

этот оператор записывается так (1-я форма оператора):

$$\begin{bmatrix} \mathbf{b}_1 \\ \mathbf{a}_1 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_-^{1'} \mathbf{F} \mathbf{X}_-^1 & \mathbf{X}_-^{1'} \mathbf{Z}^1 \\ \mathbf{Z}^{1'} \mathbf{X}_-^1 & \mathbf{Z}^{1'} \mathbf{Z}^1 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_-^{1'} \mathbf{F} \mathbf{X}_1 \\ \mathbf{Z}^{1'} \mathbf{X}_1 \end{bmatrix}, \text{ или в более "прозрачной" - 2-й}$$

форме (учитывая, что $\mathbf{X}_-^{1'c} = \mathbf{X}_-^1 - \mathbf{V}^1$):

$$\begin{bmatrix} \mathbf{b}_1 \\ \mathbf{a}_1 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_-^{1'} \mathbf{X}_-^1 - \mathbf{V}^{1'} \mathbf{V}^1 & \mathbf{X}_-^{1'} \mathbf{Z}^1 \\ \mathbf{Z}^{1'} \mathbf{X}_-^1 & \mathbf{Z}^{1'} \mathbf{Z}^1 \end{bmatrix}^{-1} \begin{bmatrix} (\mathbf{X}_-^{1'} - \mathbf{V}^{1'}) \mathbf{X}_1 \\ \mathbf{Z}^{1'} \mathbf{X}_1 \end{bmatrix}.$$

Если уравнение не идентифицировано, то обращаемая матрица в данном операторе вырождена. Если уравнение точно идентифицировано, то 2М-оценка совпадет с КМ-оценкой.

Для сверхидентифицированного уравнения можно использовать также **метод наименьшего дисперсионного отношения** (МНДО). Строгое обоснование его применимости вытекает из метода максимального правдоподобия.

Пусть \mathbf{b}^1 в уравнении $\mathbf{X}^1 \mathbf{b}^1 = \mathbf{Z}^1 \mathbf{a}_1 + \mathbf{e}_1$ оценено, и $\mathbf{X}^1 \mathbf{b}^1$ рассматривается как единая эндогенная переменная. В результате применения МНК определяются:

$$\mathbf{a}_1 = (\mathbf{Z}^{1'} \mathbf{Z}^1)^{-1} \mathbf{Z}^{1'} \mathbf{X}^1 \mathbf{b}^1,$$

$$\mathbf{e}_1 = (\mathbf{I}_N - \mathbf{Z}^1 (\mathbf{Z}^{1'} \mathbf{Z}^1)^{-1} \mathbf{Z}^{1'}) \mathbf{X}^1 \mathbf{b}^1 = (\mathbf{I}_N - \mathbf{F}^1) \mathbf{X}^1 \mathbf{b}^1,$$

$$\mathbf{e}_1' \mathbf{e}_1 = \mathbf{b}^{1'} \mathbf{X}^{1'} (\mathbf{I}_N - \mathbf{F}^1) \mathbf{X}^1 \mathbf{b}^1 = \mathbf{b}^{1'} \mathbf{W}^1 \mathbf{b}^1.$$

Теперь находится остаточная сумма квадратов при условии, что все экзогенные переменные входят в 1-е уравнение. Она равна

$$\mathbf{b}^{1'} \mathbf{W}^1 \mathbf{b}^1, \text{ где } \mathbf{W} = \mathbf{X}^{1'} (\mathbf{I}_N - \mathbf{F}) \mathbf{X}^1.$$

Тогда \mathbf{b}^1 должны были бы быть оценены так, чтобы

$$\mathbf{f} = \frac{\mathbf{b}^{1'} \mathbf{W}^1 \mathbf{b}^1}{\mathbf{b}^{1'} \mathbf{W}^1 \mathbf{b}^1} \rightarrow_{\mathbf{b}^1} \min.$$

(иначе было бы трудно понять, почему в этом уравнении присутствуют не все экзогенные переменные).

Решение этой задачи приводит к следующим условиям:

$$(\mathbf{W}^1 - \mathbf{f} \mathbf{W}) \mathbf{b}^1 = \mathbf{0},$$

из которых \mathbf{f} находится как минимальный корень соответствующего характеристического уравнения, а \mathbf{b}^1 определяется с точностью до постоянного множителя (с точностью до нормировки $\mathbf{b}_{11} = 1$).

В общем случае $\mathbf{f} > 1$, но $\mathbf{f} \rightarrow_{N \rightarrow \infty} 1$. Если данное уравнение точно идентифицировано, то $\mathbf{f} = 1$, и МНДО-оценки совпадают с КМ- и 2М-оценками.

Оператор

$$\begin{bmatrix} \mathbf{b}_1 \\ \mathbf{a}_1 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_-^{1'} \mathbf{X}_-^1 - k \mathbf{V}^{1'} \mathbf{V}^1 & \mathbf{X}_-^{1'} \mathbf{Z}^1 \\ \mathbf{Z}^{1'} \mathbf{X}_-^1 & \mathbf{Z}^{1'} \mathbf{Z}^1 \end{bmatrix}^{-1} \begin{bmatrix} (\mathbf{X}_-^{1'} - k \mathbf{V}^{1'}) \mathbf{X}_1 \\ \mathbf{Z}^{1'} \mathbf{X}_1 \end{bmatrix}$$

позволяет получить так называемые **оценки k-класса** (не путать с **k** - количеством эндогенных переменных в системе).

При **k = 0**, они являются обычными МНК-оценками для **l**-го уравнения; при **k = 1**, это - 2М-оценки; при **k = f**, - МНДО-оценки. 2М-оценки занимают промежуточное положение между МНК- и МНДО-оценками (т.к. **f > 1**). Исследования показывают, что эффективные оценки получаются при **k < 1**.

7.4. Оценка параметров всех (идентифицированных) уравнений

Из приведенной формы системы уравнений следует, что

$$\mathbf{x}'\boldsymbol{\varepsilon} = \mathbf{B}^{-1'}\mathbf{A}'\mathbf{Z}'\boldsymbol{\varepsilon} + \mathbf{B}^{-1'}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon},$$

и далее $\mathbf{E}(\mathbf{x}'\boldsymbol{\varepsilon}) = \mathbf{B}^{-1'}\mathbf{E}(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}) = \sigma^2\mathbf{B}^{-1'}\boldsymbol{\Omega}$, т.е. в общем случае все эндогенные переменные скоррелированы с ошибками во всех уравнениях. Это является основным препятствием для применения обычного МНК ко всем уравнениям по отдельности.

Но в случае, если в матрице **B** все элементы, расположенные ниже главной диагонали, равны нулю (т.е. в правой части **l**-го уравнения могут появляться только более младшие эндогенные переменные $\mathbf{x}_{l'}$, $l' < l$, и последней компонентой любого вектора \mathbf{x}^l является \mathbf{x}_l), а в матрице **Ω**, наоборот, равны нулю все элементы, расположенные выше главной диагонали или эта матрица диагональна, то $\boldsymbol{\varepsilon}_l$ не скоррелирован с переменными \mathbf{x}_l при любом **l**. Это - **рекурсивная система**, и для оценки ее параметров можно применять МНК к отдельным уравнениям.

Для оценки параметров всех идентифицированных уравнений системы можно применить **трехшаговый метод** (3М) наименьших квадратов.

Предполагается, что идентифицированы все **k** уравнений:

$$\mathbf{X}_l = \mathbf{X}_l'\boldsymbol{\beta}_l + \mathbf{Z}_l'\boldsymbol{\alpha}_l + \boldsymbol{\varepsilon}_l = \mathbf{Q}_l'\boldsymbol{\gamma}_l + \boldsymbol{\varepsilon}_l, \quad l = \overline{1, k},$$

где $\mathbf{Q}_l' = [\mathbf{X}_l', \mathbf{Z}_l']$, $\boldsymbol{\gamma}_l = \begin{bmatrix} \boldsymbol{\beta}_l \\ \boldsymbol{\alpha}_l \end{bmatrix}$.

При условии, что матрица ковариации ошибок эндогенных переменных $\sigma^2\boldsymbol{\Omega}$ одинакова во всех наблюдениях (гипотеза гомоскедастичности)

$$\mathbf{E}(\boldsymbol{\varepsilon}_l\boldsymbol{\varepsilon}_l') = \sigma^2\omega_{ll}\mathbf{I}_N, \quad \mathbf{E}(\boldsymbol{\varepsilon}_{l'}\boldsymbol{\varepsilon}_l') = \sigma^2\omega_{l'l}\mathbf{I}_N.$$

В уравнении $\mathbf{Z}'\mathbf{X}_l = \mathbf{Z}'\mathbf{Q}_l'\boldsymbol{\gamma}_l + \mathbf{Z}'\boldsymbol{\varepsilon}_l$ (*)

$\mathbf{Z}'\mathbf{X}_l$ рассматривается как вектор **n+1** наблюдений за одной эндогенной переменной, а $\mathbf{Z}'\mathbf{Q}_l'$ - как матрица **n+1** наблюдений за **n_l+k_l+1** экзогенными переменными. Поскольку матрица ковариации остатков по этому уравнению равна $\sigma^2\omega_{ll}\mathbf{Z}'\mathbf{Z}$ (т.е. отлична от $\sigma^2\mathbf{I}_N$), для получения оценок **c_l** параметров $\boldsymbol{\gamma}_l$ нужно использовать ОМНК:

$$\mathbf{c}_1 = (\mathbf{Q}^{1'} \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}' \mathbf{Q}^1)^{-1} \mathbf{Q}^{1'} \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X}_1.$$

Это - еще одна (3-я) форма записи оператора 2М-оценивания.

Первые два шага 3М совпадают с 2М, но цель их не в получении оценок \mathbf{c}_1 , а в том, чтобы оценить \mathbf{e}_1 , и затем получить оценки \mathbf{W} матрицы $\sigma^2 \Omega$:

$$\mathbf{w}_{11} = \frac{1}{N} \mathbf{e}_1' \mathbf{e}_1, \quad \mathbf{w}_{1'1} = \frac{1}{N} \mathbf{e}_1' \mathbf{e}_1.$$

Теперь все уравнения (*) записываются в единой системе:

$$\begin{bmatrix} \mathbf{Z}'\mathbf{X}_1 \\ \mathbf{Z}'\mathbf{X}_2 \\ \vdots \\ \mathbf{Z}'\mathbf{X}_k \end{bmatrix} = \begin{bmatrix} \mathbf{Z}'\mathbf{Q}^1 & \mathbf{0} & \cdot & \cdot & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}'\mathbf{Q}^2 & \cdot & \cdot & \mathbf{0} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \mathbf{0} & \mathbf{0} & \cdot & \cdot & \mathbf{Z}'\mathbf{Q}^k \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \cdot \\ \cdot \\ \gamma_k \end{bmatrix} + \begin{bmatrix} \mathbf{Z}'\boldsymbol{\varepsilon}_1 \\ \mathbf{Z}'\boldsymbol{\varepsilon}_2 \\ \cdot \\ \cdot \\ \mathbf{Z}'\boldsymbol{\varepsilon}_k \end{bmatrix} \quad (**),$$

или

$$\mathbf{Y} = \mathbf{Q}\boldsymbol{\gamma} + \boldsymbol{\eta},$$

где \mathbf{Y} - соответствующий $\mathbf{k}(\mathbf{n}+1)$ -вектор-столбец наблюдений за изучаемой переменной;

\mathbf{Q} - $\mathbf{k}(\mathbf{n}+1) \times (\sum(\mathbf{k}_1 + \mathbf{n}_1) + \mathbf{k})$ -матрица наблюдений за экзогенными переменными;

$\boldsymbol{\gamma}$ - $(\sum(\mathbf{k}_1 + \mathbf{n}_1) + \mathbf{k})$ -вектор-столбец параметров регрессии;

$\boldsymbol{\eta}$ - $\mathbf{k}(\mathbf{n}+1)$ -вектор столбец остатков по наблюдениям.

Легко проверить, что матрица ковариации остатков $\boldsymbol{\eta}$ удовлетворяет следующему соотношению:

$$\mathbf{E}(\boldsymbol{\eta}\boldsymbol{\eta}') = \sigma^2 \Omega \otimes (\mathbf{Z}'\mathbf{Z}),$$

где \otimes - операция прямого умножения матриц.

Для нее имеется оценка: $\mathbf{k}(\mathbf{n}+1) \times \mathbf{k}(\mathbf{n}+1)$ -матрица $\boldsymbol{\Sigma} = \mathbf{W} \otimes (\mathbf{Z}'\mathbf{Z})$.

Эта матрица отлична от $\sigma^2 \mathbf{I}_{\mathbf{k}(\mathbf{n}+1)}$, поэтому на 3-м шаге 3М-оценивания к единой системе (**) применяется ОМНК и получается окончательная оценка \mathbf{c} параметров $\boldsymbol{\gamma}$:

$$\mathbf{c} = (\mathbf{Q}' \boldsymbol{\Sigma}^{-1} \mathbf{Q})^{-1} \mathbf{Q}' \boldsymbol{\Sigma}^{-1} \mathbf{Y}$$

В таком виде оператор 3М-оценивания используется для всех сверхидентифицированных уравнений. Для точно идентифицированных уравнений он имеет более сложную форму. Но для таких уравнений всегда можно применить КМ-оценивание.

Теоретические вопросы и задания

1(*). Доказать, что параметры уравнений неодновременной системы можно оценивать обычным МНК, даже если ошибки эндогенных переменных скоррелированы.

2. Вывести приведенные уравнения оценки параметров неодновременной системы для случая априорных ограничений на эти параметры.

3(**). Доказать сформулированное необходимое и достаточное условие идентификации уравнения одновременной системы.

4(*). Показать, что обычный МНК, примененный к системе одновременных уравнений, дает в общем случае смещенные оценки.

5(*). Убедиться в том, что все три приведенные формы оператора 2М-оценивания эквивалентны.

6(*). Доказать, что в случае точной идентификации уравнения 2М- и КМ-оценки его параметров одинаковы, а в случае, если уравнение не идентифицировано, то матрица, обратная к которой фигурирует в операторе 2М-оценивания, вырождена.

7(**). Вывести указанные условия для нахождения МНДО-оценок.

8(*). Показать, что в случае точной идентификации уравнения МНДО-, 2М- и КМ-оценки совпадают.