МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ОДЕСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ І. І. МЕЧНИКОВА
ІНСТИТУТ МАТЕМАТИКИ, ЕКОНОМІКИ І МЕХАНІКИ
Кафедра оптимального керування і економічної кібернетики

МЕТОДИЧНІ ВКАЗІВКИ

для самостійної роботи з дисципліни «Економетрика»

ДЛЯ СТУДЕНТІВ НАПРЯМІВ ПІДГОТОВКИ
6.040201 «МАТЕМАТИКА», 6.040301 «ПРИКЛАДНА МАТЕМАТИКА»,
6.030501 «ЕКОНОМІЧНА ТЕОРІЯ», 6.030203 «МІЖНАРОДНІ ЕКОНОМІЧНІ ВІДНОСИНИ», 6.030601 «МЕНЕДЖМЕНТ»

Частина А

Одеса
Демидов
2014

УДК 330.43
ББК 65в6

Методичні вказівки для самостійної роботи з дисципліни «Економетрика» для студентів напрямів підготовки 6.040201 «математика», 6.040301 «прикладна математика», 6.030501 «економічна теорія», 6.030203 «міжнародні економічні відносини», 6.030601 «менеджмент».

Методичні вказівки для самостійної роботи з дисципліни «Економетрика» допоможуть майбутнім фахівцям економістам і математикам опанувати методику побудови і дослідження економетричних моделей, що описують соціально-економічні явища.

Автори-укладачі: **Яровий А. Т.,** кандидат фізико-математичних наук, доцент кафедри оптимального керування і економічної кібернетики;
Страхов Є. М., кандидат фізико-математичних наук, старший викладач кафедри оптимального керування і економічної кібернетики.

Рецензенти: **Мацкул В. М.,** кандидат фізико-математичних наук, доцент кафедри математичних методів аналізу економіки Одеського національного економічного університету;
Васильєв О. Б., кандидат фізико-математичних наук, доцент кафедри оптимального керування і економічної кібернетики.

Рекомендовано до друку Вченою радою ІМЕМ ОНУ імені І. І. Мечникова (протокол № 5 від 10.06.2014).

ВСТУП

Діяльність у довільній галузі економіки вимагає від фахівця знання сучасних методів дослідження економетричних взаємозв'язків, що ґрунтуються на економетричних моделях. Побудова і дослідження економетричних моделей – центральна тема дисципліни «Економетрика». Вона об'єднує сукупність теоретичних результатів, методів і моделей та на основі економічної теорії, економічної статистики надає конкретний кількісний вираз закономірностям, що обумовлені економічною теорією.

У цьому посібнику викладемо деякі з основних положень побудови і дослідження економетричних моделей у випадку виконання передумов використання методу 1МНК, а також при наявності автокореляції і гетероскедастичності залишків, мультиколінеарності регресорів. Крім методів побудови і дослідження моделей у посібнику розглядаються приклади їх застосування, що допоможе студентам краще опанувати основні положення дисципліни.

РОЗДІЛ 1. ПОБУДОВА БАГАТОФАКТОРНОЇ РЕГРЕСІЙНОЇ МОДЕЛІ ПРИ ВИКОНАННІ ВСІХ ПЕРЕДУМОВ 1МНК

ТЕМА 1. КОЕФІЦІЄНТ КОРЕЛЯЦІЇ

При побудові економічних моделей виникає питання: які ж чинники включати до моделі? Зрозуміло, що ті чинники, які найбільш тісно пов'язані з результуючим показником. І тому необхідно мати правила, за допомогою яких можна визначати напрямок і тісноту зв'язку між результуючим показником і чинником. Для цього використовують парний коефіцієнт кореляції і частковий коефіцієнт кореляції.

Коефіцієнт кореляції

Коефіцієнт кореляції є одним з основних показників взаємозалежності випадкових величин. Його також називають парним коефіцієнтом кореляції або лінійним коефіцієнтом кореляції.

Отже, парний коефіцієнт кореляції характеризує тісноту і напрямок зв'язку між двома корелюючими ознаками у випадку наявності між ними лінійної залежності.

Ці дві ознаки повинні вести себе як двовимірна нормальна випадкова величина.

Вибіркова величина парного коефіцієнта кореляції визначається за формулою

$$\hat{r} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right] \cdot \left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}, \quad (1)$$

де n – довжина вибірки, а x і y – ознаки.

Розглянемо властивості парного коефіцієнта кореляції.

1. Значення парного коефіцієнта кореляції належать проміжку $[-1; 1]$.

Якщо $\hat{r} > 0$, то зв'язок між x і y прямий (обидві зростають або обидві зменшуються одночасно), якщо ж $\hat{r} < 0$, то зв'язок – обернений (якщо одна ознака зростає, то друга – спадає і навпаки).

2. Якщо випадкові величини x і y статистично незалежні, то $\hat{r} = 0$.
3. Із того, що $\hat{r} = 0$, випливає некорельованість випадкових величин.
4. Якщо $|\hat{r}| = 1$, то це означає, що між змінними x і y існує функціональний лінійний зв'язок і навпаки: якщо між x і y існує функціональна лінійна залежність, то $|\hat{r}| = 1$.

Вибірковий коефіцієнт кореляції, що визначений формулою (1), можна розраховувати для довільної двовимірної системи, яка має спільний нормальний розподіл випадкових величин.

Коефіцієнт кореляції разом з середніми і дисперсіями випадкових величин складає ті п'ять параметрів, що дають вичерпні відомості про стохастичну залежність величин, тому що однозначно визначають їх двовимірний закон розподілу.

У випадках, коли розподіл відхиляється від нормального, одна із величин не підпорядкована нормальному законові розподілу, величини не є випадковими, коефіцієнт кореляції можна використовувати лише як одну з можливих характеристик ступеня тісноти зв'язку. При цьому, не дивлячись на те, що в загальному випадку поки що не запропоновано характеристики лінійного зв'язку, яка мала б очевидні переваги в порівнянні з r, його інтерпретація досить часто є ненадійною. Якщо апріорі допускається можливість відхилення від лінійного вигляду залежності, то можна побудувати приклади, коли при $r = 0$ існує чисто функціональна залежність між ознаками. Тому про величини, для яких $r = 0$, кажуть, що вони некорельовані, і тільки після професійного аналізу можна сказати, чи є вони незалежними. І, навпаки, з високого ступеня корельованості величин при великих відхиленнях їх розподілів від нормального ще не виходить їх досить тісна залежність.

Тест на перевірку значущості коефіцієнта кореляції

Після оцінки лінійного коефіцієнта кореляції виникає питання: яку величину вибіркового коефіцієнта кореляції можна вважати достатньою для статистично обґрунтованого висновку про наявність кореляційного зв'язку між змінними? Надійність статистичних характеристик, в тому числі і $\hat{r}$, зменшується зі зменшенням об'єму вибірки, а тому можливі випадки, коли відхилення від нуля отриманої величини вибіркового коефіцієнта кореляції $\hat{r}$ є статистично незначущим, тобто цілком обумовленим випадковими коливаннями вибірки, за якою він розрахований. Відповісти на це питання допомагає знання закону імовірнісного розподілу $\hat{r}$. У випадку спільної нормальної розподіленості змінних і при достатньо великому об'ємові вибірки n розподіл $\hat{r}$ можна вважати наближено нормальним з середнім, що дорівнює своєму теоретичному значенню r. Однак необхідно враховувати, що при малих значеннях n і r, близьких до ± 1, це наближення є досить грубим. Крім того, при малих n слід приймати до уваги, що величина $\hat{r}$ є зміщеною оцінкою свого теоретичного значення r.

Відносно добрий ступінь наближення нормального розподілу при малих значеннях $|r|$ дозволяє отримати простий критерій перевірки гіпотези $r = 0$, тобто гіпотези про відсутність кореляційного зв'язку між змінними. Тест на перевірку значущості коефіцієнта кореляції має такий алгоритм:

1) формулювання гіпотез: $H_A : r \neq 0,\ H_0 : r = 0$;
2) задаємо рівень значущості α;
3) розраховуємо $t_{cm} = \left|\hat{r}\right| \sqrt{\dfrac{n-2}{1-\hat{r}^2}}$;

4) за таблицею Ст'юдента визначаємо t_{kp} : $t_{kp} = t\left(\frac{\alpha}{2}; n-2\right)$;

5) якщо $t_{cm} > t_{kp}$, то гіпотеза H_0 відхиляється з імовірністю $(1-\alpha)\cdot 100\%$, а це означає, що приймається гіпотеза H_A, тобто вибірковий коефіцієнт кореляції значимо відрізняється від нуля з імовірністю $(1-\alpha)\cdot 100\%$.

Часткові коефіцієнти кореляції

Часткові коефіцієнти кореляції характеризують ступінь тісноти зв'язку між двома ознаками при умові, що всі інші фіксовані на певному рівні, тобто оцінюється зв'язок між ознаками в «чистому» вигляді.

Існує дві взаємопов'язані обставини, що перешкоджають широкому практичному використанню часткових характеристик статистичного зв'язку в загальному (негаусівському) випадку:

а) часткові характеристики статистичного зв'язку залежать від рівнів x заважаючих змінних;

б) для підрахунку вибіркових значень часткових характеристик статистичного зв'язку необхідно мати вибірку спеціальної структури, яка забезпечувала б наявність хоча би декількох спостережень при кожному з заданого ряду фіксованих значень x заважаючих змінних.

Але, якщо випадкові змінні підпорядковуються багатовимірному нормальному законові, то згадані неподобства зникають, так як в цьому випадку часткові коефіцієнти кореляції не залежать від рівня заважаючих змінних x.

Має місце така оцінка часткового коефіцієнта кореляції

$$\hat{r}_{ij|\cdot} = \frac{-R_{ij}}{\left(R_{ii}\cdot R_{jj}\right)^{1/2}}, \qquad (2)$$

де $\hat{r}_{ij|\cdot}$ означає частковий коефіцієнт кореляції між змінними x_i і x_j при фіксованих значеннях всіх інших змінних, а R_{kl} – алгебраїчне доповнення для елемента $\hat{r}_{kl}$ в детермінанті кореляційної матриці R ознак.

Можна користуватися і такою формулою

$$\hat{r}_{ij|\cdot} = \frac{-z_{ij}}{\left(z_{ii}\cdot z_{jj}\right)^{1/2}}, \qquad (3)$$

де z_{kl} – елементи матриці Z, що є оберненою до R.

У випадку залежності y від двох чинників x_1 і x_2 часткові коефіцієнти кореляції оцінюються наступним чином:

$$\hat{r}_{yx_1|x_2} = \frac{\hat{r}_{yx_1} - \hat{r}_{yx_2} \cdot \hat{r}_{x_1x_2}}{\sqrt{\left(1 - \hat{r}^2_{yx_2}\right)\left(1 - \hat{r}^2_{x_1x_2}\right)}},$$

$$\hat{r}_{yx_2|x_1} = \frac{\hat{r}_{yx_2} - \hat{r}_{yx_1} \cdot \hat{r}_{x_1x_2}}{\sqrt{\left(1 - \hat{r}^2_{yx_1}\right)\left(1 - \hat{r}^2_{x_1x_2}\right)}}, \quad (4)$$

де $\hat{r}_{yx_1}$, $\hat{r}_{yx_2}$, $\hat{r}_{x_1x_2}$ – оцінки парних коефіцієнтів кореляції.

Практика довела, що часткові коефіцієнти кореляції, що визначені формулами (2), (3) і (4), є, як правило, задовільним вимірювачами очищеного лінійного зв'язку між x_i і x_j при фіксованих значеннях інших змінних і у випадку, коли розподіл показників $y, x_1, ..., x_p$ відрізняється від нормального. Ці показники зв'язку можна інтерпретувати як показники тісноти очищеного зв'язку, що усереднені по можливим значенням фіксованих на певному рівні «заважаючих» змінних.

Тест на значущість часткового коефіцієнта кореляції

При дослідженні статистичних властивостей вибіркового часткового коефіцієнта кореляції порядку k (тобто при виключенні опосередкованого впливу k «заважаючих» змінних) необхідно скористатися тим, що він розподілений так, як і парний вибірковий коефіцієнт кореляції між тими ж змінними, тільки скрізь необхідно замінити об'єм вибірки n на $n - k$.

Отже, t_{cm} розраховується за формулою

$$t_{cm} = \left|\hat{r}_{yx|\cdot}\right| \sqrt{\frac{n-k-2}{1 - \hat{r}^2_{yx|\cdot}}},$$

а $t_{kp} = t\left(\frac{\alpha}{2}; n-k-2\right)$.

Якщо $t_{cm} > t_{kp}$, то з імовірністю $(1-\alpha)\cdot 100\%$ частковий коефіцієнт кореляції значимо відрізняється від нуля.

Приклад 1. Розглянемо показники виробничо-господарської діяльності 25 підприємств машинобудування: y – продуктивність праці, x_2 – трудомісткість одиниці продукції, x_3 – середньорічний фонд заробітної платні підприємства, x_4 – фондоозброєність праці, x_5 – невиробничі витрати. Вони мають такий вигляд:

Номер підприємства	Y	X2	X3	X4	X5
1	9,26	0,23	47750	6,4	17,72
2	9,38	0,24	50391	7,8	18,39
3	12,11	0,19	43149	9,76	26,46
4	10,81	0,17	41089	7,9	22,37
5	9,35	0,23	14257	5,35	28,13
6	9,87	0,43	22661	9,9	17,55
7	8,17	0,31	52509	4,5	21,92
8	9,12	0,26	14903	4,88	19,52
9	5,88	0,49	25587	3,46	23,99
10	6,3	0,36	16821	3,6	21,76
11	6,22	0,37	19459	3,56	25,68
12	5,49	0,43	12973	5,65	18,13
13	6,5	0,35	50907	4,28	25,74
14	6,61	0,38	6920	8,85	21,21
15	4,32	0,42	5736	8,52	22,97
16	7,37	0,3	26705	7,19	16,38
17	7,02	0,32	20068	4,82	13,21
18	8,25	0,25	11487	5,46	14,48
19	8,15	0,31	32029	6,2	13,38
20	8,72	0,26	18946	4,25	13,69
21	6,64	0,37	28025	5,38	16,66
22	8,1	0,29	20968	5,88	15,06
23	5,52	0,34	11049	9,27	20,09
24	9,37	0,23	45893	4,36	15,98
25	13,17	0,17	99400	10,31	18,27

Необхідно розрахувати парні і часткові коефіцієнти кореляції між y та x_i, а також перевірити їх на значущість.

Скориставшись формулою (1), отримаємо:

$$r_{yx_2} = -0.801,\ r_{yx_3} = 0.664,\ r_{yx_4} = 0.400,\ r_{yx_5} = -0.053.$$

Розрахуємо t-статистики. Маємо

$$t_{cm\,yx_2} = 6.412,\ t_{cm\,yx_3} = 4.262,\ t_{cm\,yx_4} = 2.091,\ t_{cm\,yx_5} = 0.253.$$

Покладемо $\alpha = 0.1$, тоді $t_{kp} = t\left(\frac{\alpha}{2}; n-2\right) = t\left(0.05; 23\right) = 1.714$.

Остаточно маємо, що всі парні коефіцієнти кореляції, крім останнього, значимо відрізняються від нуля з імовірністю 90%.

Тепер розрахуємо часткові коефіцієнти кореляції. Маємо

$$r_{yx_2|\cdot} = -0.718,\ r_{yx_3|\cdot} = 0.467,\ r_{yx_4|\cdot} = 0.373,\ r_{yx_5|\cdot} = -0.0004.$$

Їх статистики такі:

$$t_{cm\,yx_2|\cdot} = 4.837,\ t_{cm\,yx_3|\cdot} = 2.479,\ t_{cm\,yx_4|\cdot} = 1.884,\ t_{cm\,yx_5|\cdot} = 0.002.$$

Далі, $t_{kp} = t\left(\frac{\alpha}{2}; n-k-2\right) = t\left(0.05; 20\right) = 1.725$.

Так як $t_{cm\,yx_i|\cdot}\left(i = 2,3,4\right) > t_{kp}$, то перші три часткові коефіцієнти кореляції значимо відрізняються від нуля з імовірністю 90%, а останній – ні.

Зазначимо, що часткові коефіцієнти кореляції трохи менші за абсолютною величиною, ніж парні. А це означає, що існує невеликий зв'язок між x_i.

Дивлячись на величини парних і часткових коефіцієнтів кореляції, можна зробити висновок, що до економічної моделі, що показує зв'язок між y і x_i, не доцільно включати чинник x_5. Найбільш впливовим чинником на продуктивність праці є трудомісткість одиниці продукції, потім – середньорічний фонд заробітної платні підприємства і потім фондоозброєність праці. Невиробничі витрати майже не впливають на продуктивність праці.

ТЕМА 2. ЛІНІЙНА БАГАТОФАКТОРНА МОДЕЛЬ. ОСНОВНІ ПРИПУЩЕННЯ У БАГАТОФАКТОРНОМУ РЕГРЕСІЙНОМУ АНАЛІЗІ. ОЦІНКА ПАРАМЕТРІВ БАГАТОФАКТОРНОЇ І ПАРНОЇ РЕГРЕСІЇ

Узагальнену багатофакторну лінійну регресійну модель будемо записувати у вигляді

$$y = \sum_{i=1}^{N} b_i x_i + \varepsilon, \tag{1}$$

де y – залежна (ендогенна) змінна, $x_1, x_2, ..., x_N$ – незалежні (екзогенні) змінні, $b_1, b_2, ..., b_N$ – параметри моделі, які необхідно оцінити, ε – неспостережувана випадкова величина.

Зазначимо, що узагальнена регресійна модель – це модель, що дійсна для всієї генеральної сукупності. Так як випадкова величина ε – неспостережувана, то можна тільки робити припущення відповідно до закону її розподілу.

На відміну від узагальненої регресійної моделі, вибіркова модель будується для певної вибірки. Невідомі параметри вибіркової моделі є випадковими величинами, математичне очікування яких дорівнює параметрам узагальненої моделі, а випадкові величини можна оцінити, виходячи з вибіркових даних.

Будемо вважати, що вибіркова лінійна багатофакторна модель має вигляд

$$y = \sum_{i=1}^{N} \beta_i x_i + u, \tag{2}$$

де y – задана залежна змінна, x_i, $i = \overline{1, N}$ – задані незалежні змінні, u – випадкова величина (помилка).

Будемо вважати модель лінійною за параметрами β_i.

Модель (2) можна записати у вигляді

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_N x_{Nt} + u_t, \quad t = \overline{1, T}. \tag{3}$$

Зазначимо, що $x_{1t} \equiv 1$, $t = \overline{1, T}$. T – довжина вибірки, кількість спостережень.

Можна ще модель записати і у матричному вигляді

$$y = X\beta + u,$$

де $y = (y_1, ..., y_T)^T$ – вектор-стовпець значень y_t;

$\beta = (\beta_1, ..., \beta_N)^T$ – вектор-стовпець параметрів;

$u = (u_1, ..., u_T)^T$ – вектор-стовпець випадкової змінної u;

$$X = \begin{pmatrix} 1 & x_{21} & \cdots & x_{N1} \\ 1 & x_{22} & \cdots & x_{N2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{2T} & \cdots & x_{NT} \end{pmatrix}$$ – матриця значень змінних $x_1, ..., x_N$.

Через $D = \left[Y \middle| X \right]$ позначимо матрицю початкових даних, де Y і X визначені вище.

У регресійному аналізі елементи моделі (1) повинні задовольняти наступним умовам:

У1. Відсутність систематичних помилок спостережень y_t, тобто математичне очікування u_t дорівнює нулеві, а також $M(u) = 0$. Інколи випадковий член буде додатнім, інколи від'ємним, але він не повинен мати систематичного зміщення в жодному з двох положень.

Фактично, якщо рівняння регресії включає постійний член, то розумно вважати, що ця умова виконується автоматично, так як роль константи полягає у визначенні довільної систематичної тенденції в y, яку не враховують пояснюючі змінні, що включені до моделі.

У2. Дисперсійно-коваріаційна матриця помилок u має вигляд: $\sum_u = \sigma_u^2 I$, де σ_u^2 – дисперсія залишків, а I – одинична матриця. Ця умова стверджує, що помилки мають постійну дисперсію (гомоскедастичні) і вільні від кореляції, тобто

$$\sigma_{u_i u_j} = \begin{cases} \sigma_u^2, & i = j, \\ 0, & i \neq j, \ i, j = \overline{1, T}. \end{cases}$$

У3. Залишки u нормально розподілені: $U \sim N(0, \sigma_u^2 I)$.

Якщо випадковий член u нормально розподілений, то так же будуть розподілені і коефіцієнти регресії, а це дасть нам можливість робити тести і розраховувати інтервали довіри.

У4. Екзогенні змінні виміряються без помилок і утворюють лінійно-незалежні вектори, тобто $rang\, X = N$.

У5. Змінні $x_i, \ i = \overline{1, N}$ не корелюють з помилками u.

Якщо виконуються згадані припущення, то для оцінки невідомих параметрів $\beta_i, \ i = \overline{1, N}$ можна застосувати метод найменших квадратів (1МНК). Це означає, що β будемо знаходити з умови:

$$\min_{\beta} \sum_{i=1}^{T} u_i^2,$$

тобто

$$\widehat{\beta} = \arg\min_{\beta} \left[\left(y - X\beta \right)^T \left(y - X\beta \right) \right] \qquad (3)$$

($\widehat{\boldsymbol{\beta}}$ означає оцінку параметра $\boldsymbol{\beta}$ методом 1МНК). З (3) отримаємо

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}. \tag{4}$$

Підкладаючи $\widehat{\boldsymbol{\beta}}_k$ $(\boldsymbol{k} = \overline{\boldsymbol{1,N}})$ у регресійне рівняння (2), отримаємо оцінену за допомогою 1МНК емпіричну регресійну функцію

$$\hat{\boldsymbol{y}}_t = \sum_{i=1}^{N} \widehat{\boldsymbol{\beta}}_i \boldsymbol{x}_i \text{ або } \widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}. \tag{5}$$

Емпіричній коефіцієнт $\widehat{\boldsymbol{\beta}}_k$ має таку інтерпретацію: зміна величини $\boldsymbol{k}$-го регресора (незалежної змінної $\boldsymbol{x}_k$) на одиницю при інших рівних умовах викликає зміну величини $\hat{\boldsymbol{y}}$ на $\widehat{\boldsymbol{\beta}}_k$ одиниць.

Якщо ж регресія парна: $\boldsymbol{y} = \beta_1 x_1 + \beta_2 x_2 + \boldsymbol{u}$, то для оцінки параметрів β_1 і β_2 можна користуватися формулами:

$$\widehat{\beta}_2 = \frac{\sum_{t=1}^{T} x_{2t}y_t - \frac{1}{T}\sum_{t=1}^{T} x_{2t}\sum_{t=1}^{T} y_t}{\sum_{t=1}^{T} x_{2t}^2 - \frac{1}{T}\left(\sum_{t=1}^{T} x_{2t}\right)^2}, \quad \widehat{\beta}_1 = \bar{y} - \widehat{\beta}_2\bar{x}_2, \text{ де } \bar{y} = \frac{\sum_{t=1}^{T} y_t}{T}, \quad \bar{x}_2 = \frac{\sum_{t=1}^{T} x_{2t}}{T}.$$

Далі оцінюємо дисперсію залишків таким чином:

$$\hat{\sigma}_u^2 = \frac{\sum_{t=1}^{T} \hat{u}_t^2}{T-N}, \quad \hat{u}_t = y_t - \hat{y}_t, \quad t = \overline{1,T}.$$

Для порівняння моделей використовується величина

$$MSE = \frac{1}{T}\sum_{t=1}^{T} \hat{u}_t^2 \text{ – середній квадрат модельної похибки.}$$

Якщо порівнюються декілька моделей, то найкращою вважається модель з найменшими значеннями $\hat{\sigma}_u^2$ і MSE.

Для оцінки однієї моделі використовується величина

$$MAPE = \frac{100\%}{T}\sum_{t=1}^{T} \frac{\left|\hat{u}_t\right|}{y_t},$$

коефіцієнт апроксимації або середня відносна величина модельної помилки. Ця величина має такі порогові значення:

Оцінка MAPE	Характеристика якості регресійної моделі
< 10	висока точність
10 ÷ 20	добра точність
20 ÷ 50	задовільна точність
> 50	незадовільна точність

Приклад 2. За даними прикладу 1 досліджується залежність продуктивності праці від трудомісткості одиниці продукції, середньорічного фонду заробітної платні підприємства, фондоозброєності праці і невиробничих витрат.

Необхідно:

1. Побудувати моделі, що виражають:
   А: залежність продуктивності праці від трудомісткості одиниці продукції, середньорічного фонду заробітної платні підприємства, фондоозброєності і невиробничих витрат;
   Б: залежність продуктивності праці від трудомісткості одиниці продукції;
   В: залежність продуктивності праці від середньорічного фонду заробітної платні підприємства;
   Г: залежність продуктивності праці від фондоозброєності праці;
   Д: залежність продуктивності праці від невиробничих витрат.
2. Для кожної з моделей оцінити $\hat{\sigma}_u^2$, MSE, $MAPE$, $M\left(\hat{u}\right)$, де
$$M\left(\hat{u}\right) = \frac{1}{T}\sum_{t=1}^{T} u_t .$$
3. Зробити порівняльну характеристику моделей і визначити найбільш впливовий чинник.

Отримані розрахунки занесемо в таблицю.

Модель		$\hat{\sigma}_u^2$	MSE	MAPE	$M\left(\hat{u}\right)$
А	$\hat{y} = 10.439\,x_1 - 14.755\,x_2 + +0.00003\,x_3 + 0.198\,x_4 - 0.00009\,x_5$	1.280	1.024	9.4%	$-9\cdot10^{-16}$
Б	$\hat{y} = 14.156\,x_1 - 19.767\,x_2$	1.682	1.547	10.6%	$4\cdot10^{-15}$
В	$\hat{y} = 6.056\,x_1 + 0.00007\,x_3$	2.622	2.412	17%	$-9\cdot10^{-16}$
Г	$\hat{y} = 5.599\,x_1 + 0.392\,x_4$	3.942	3.627	21.2%	$-6\cdot10^{-15}$
Д	$\hat{y} = 8.571\,x_1 + 0.026\,x_5$	4.679	4.305	22.2%	$-2\cdot10^{-15}$

Висновки

1. Перше припущення ($M(\hat{u}) = 0$) виконується у всіх п'яти моделях.
2. За показниками $\hat{\sigma}_u^2$ і MSE найкращою моделлю є модель А, далі моделі Б, В, Г і Д.
3. Так як серед парних регресій найкращою є модель Б, то трудомісткість одиниці продукції є найбільш впливовим чинником.
4. Коефіцієнт апроксимації так характеризує моделі: у моделі А відмінна апроксимація, Б і В – добра, а у Г і Д – задовільна.

ТЕМА 3. СТАНДАРТИЗОВАНІ КОЕФІЦІЄНТИ РЕГРЕСІЇ. КОЕФІЦІЄНТИ ЕЛАСТИЧНОСТІ. КОВАРІАЦІЙНА МАТРИЦЯ ДЛЯ $\widehat{\boldsymbol{\beta}}$. СТАТИСТИЧНІ ВЛАСТИВОСТІ 1МНК-ОЦІННИКА $\widehat{\boldsymbol{\beta}}$

Нагадаємо, що коефіцієнт $\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}$ має таку інтерпретацію: якщо змінити величину $\boldsymbol{k}$-го регресора на одну одиницю при інших рівних умовах, то $\widehat{\boldsymbol{y}}$ зміниться на $\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}$ одиниць. Але звідси не випливає, що чим більший за модулем коефіцієнт при регресорі, тим впливовіший цей регресор. Це пов'язано з тим, що регресори мають різні виміри. Однак, такий висновок можна робити, якщо перейти до стандартизованого рівняння. Для цього початкові дані необхідно стандартизувати за правилом:

$$\boldsymbol{z_k = \frac{z_k - \bar{z}}{\widehat{\sigma}_z}}, \text{ де } \boldsymbol{\bar{z} = \frac{\sum_{i=1}^{T} z_i}{T}}, \ \boldsymbol{\widehat{\sigma}_z = \sqrt{\frac{\sum_{i=1}^{T}\left(z_i - \bar{z}\right)^2}{T-1}}}.$$

Потім до стандартизованих (нормалізованих) початкових даних застосовується метод 1МНК, і ми отримуємо стандартизоване рівняння регресії.

Можна піти іншим шляхом. Якщо ми вже отримали регресійне рівняння, то оцінені значення стандартизованих регресійних коефіцієнтів можна обчислити за допомогою такої формули:

$\boldsymbol{\widehat{\beta}_k^{cm} = \widehat{\beta}_k \frac{\widehat{\sigma}_{x_k}}{\widehat{\sigma}_y}}$, де $\widehat{\boldsymbol{\sigma}}_{\boldsymbol{x_k}}$ і $\widehat{\boldsymbol{\sigma}}_{\boldsymbol{y}}$ – стандартні відхилення:

$$\boldsymbol{\widehat{\sigma}_{x_k} = \sqrt{\frac{\sum_{i=1}^{T}\left(x_{ki} - \bar{x}_k\right)^2}{T-1}}}, \quad \boldsymbol{\widehat{\sigma}_y = \sqrt{\frac{\sum_{i=1}^{T}\left(y_i - \bar{y}\right)^2}{T-1}}}, \quad \boldsymbol{k = \overline{2, N}}. \qquad (1)$$

Стандартизований регресійний коефіцієнт $\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}^{\boldsymbol{cm}}$ вказує на те, як за великий при інших однакових умовах оцінений типовий ефект впливу $\boldsymbol{k}$-го регресора порівняно з типовим ефектом зміни регресанда.

Чим більший за абсолютною величиною оцінений стандартизований коефіцієнт, тим більш впливовим є регресор.

Зазначимо, що $\widehat{\boldsymbol{\sigma}}_{\boldsymbol{x_1}} = \boldsymbol{0}$, і тому $\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}^{\boldsymbol{cm}} = \boldsymbol{0}$.

Приклад 3. Для моделі А:

$$\boldsymbol{\widehat{y} = 10.439\, x_1 - 14.755\, x_2 + 0.00003\, x_3 + 0.198\, x_4 - 0.00009\, x_5}$$

розрахувати стандартизовані коефіцієнти і визначити найбільш впливовий регресор.

Розрахуємо середні значення регресорів і середньоквадратичні відхилення (стандартні відхилення):

$$\boldsymbol{\bar{x}_2 = 0.308, \ \bar{x}_3 = 29587.28, \ \bar{x}_4 = 6.301, \ \bar{x}_5 = 19.550,}$$

$$\widehat{\sigma}_{x_1} = 0,\ \widehat{\sigma}_{x_2} = 0.086,\ \widehat{\sigma}_{x_3} = 20719.923,\ \widehat{\sigma}_{x_4} = 2.163,\ \widehat{\sigma}_{x_5} = 4.340.$$

Тоді за формулою (1) отримаємо:

$$\widehat{\beta}_1^{cm} = 0,\ \widehat{\beta}_2^{cm} = -0.598,\ \widehat{\beta}_3^{cm} = 0.306,\ \widehat{\beta}_4^{cm} = 0.202,\ \widehat{\beta}_5^{cm} = -0.0002.$$

Остаточно маємо:

$$\widehat{y}^{cm} = -0.598\,x_2 + 0.306\,x_3 + 0.202\,x_4 - 0.0002\,x_5.$$

Так як $\left|\widehat{\beta}_2^{cm}\right| > \left|\widehat{\beta}_3^{cm}\right| > \left|\widehat{\beta}_4^{cm}\right| > \left|\widehat{\beta}_5^{cm}\right|$, то найвпливовішим чинником на продуктивність праці є трудомісткість одиниці продукції, потім середньорічний фонд заробітної платні підприємства, далі фондоозброєність праці і невиробничі витрати. Такий же результат ми отримали раніше, коли порівнювали моделі між собою.

Коефіцієнт еластичності

Коли ми інтерпретували регресійні коефіцієнти, то приймали до уваги одиниці виміру регресорів і регресанда. Для визначення міри впливу регресора на регресанд без урахування одиниць їх виміру використовують коефіцієнт еластичності.

Коефіцієнт еластичності показує, на скільки відсотків зміниться регресанд, якщо при інших рівних умовах k-й регресор збільшити на один відсоток.

Оцінена еластичність регресанда y відносно регресора x_k розраховується за формулою:

$$\widehat{\varepsilon}_k = \widehat{\beta}_k \frac{x_k^*}{y^*} \quad \left(y^* \neq 0,\ k = \overline{2, N}\right), \tag{2}$$

де $x_k^*,\ y^*$ – значення k-го регресора і регресанда, що визначають точку регресійної функції, для якої розраховується коефіцієнт еластичності. Частіше за все використовують значення арифметичних середніх $\overline{x}_k,\ \overline{y}$ у базовому часовому або просторовому ряді.

Приклад 4. Для моделі А розрахувати коефіцієнти еластичності $\widehat{\varepsilon}_2,\ \widehat{\varepsilon}_3,\ \widehat{\varepsilon}_4,\ \widehat{\varepsilon}_5$. Дати інтерпретацію отриманим результатам.

Користуючись формулою (2), розрахуємо коефіцієнти еластичності при

$$\overline{x}_2 = 0.308,\ \overline{x}_3 = 29587.28,\ \overline{x}_4 = 6.301,\ \overline{x}_5 = 19.550,\ \overline{y} = 8.068.$$

Маємо:

$$\hat{\varepsilon}_2 = \hat{\beta}_2 \frac{\overline{x_2}}{\underline{y}} = -14.755 \cdot \frac{0.308}{8.068} = -0.563,$$
$$\hat{\varepsilon}_3 = \hat{\beta}_3 \frac{\overline{x_3}}{\underline{y}} = 0.00003 \cdot \frac{29587.28}{8.068} = 0.115,$$
$$\hat{\varepsilon}_4 = \hat{\beta}_4 \frac{\overline{x_4}}{\underline{y}} = 0.198 \cdot \frac{6.301}{8.068} = 0.155,$$
$$\hat{\varepsilon}_5 = \hat{\beta}_5 \frac{\overline{x_5}}{\underline{y}} = -0.00009 \cdot \frac{19550}{8.068} = -0.0002.$$

Наприклад, якщо трудомісткість одиниці продукції збільшити на один відсоток, то продуктивність праці зменшиться на 0.563%.

Дисперсійно-коваріаційна матриця

У класичній регресійній моделі вектор залишків U і залежний від нього Y є випадковими змінними. А так як вектор Y у свою чергу входить до функції оцінювання коефіцієнтів регресії методом 1МНК, то ці коефіцієнти також є випадковими. Для характеристики випадкових змінних $\hat{\beta}_k \left(k = \overline{1,N}\right)$ разом з математичним очікуванням використовується дисперсія $\sigma^2_{\beta_k}$ і коваріація $\sigma_{\beta_i\beta_k} \left(k,i = \overline{1,N},\ k \neq i\right)$. Значення цих параметрів класичної регресійної моделі утворюють дисперсійно-коваріаційну матрицю $\Sigma_{\hat{\beta}}$ вимірності $N \times N$. Оцінена методом 1МНК дисперсійно-коваріаційна матриця має вигляд

$$\hat{\Sigma}_{\hat{\beta}} = \hat{\sigma}^2_u \left(X^T X\right)^{-1} = \begin{pmatrix} \hat{\sigma}^2_{\hat{\beta}_1} & \hat{\sigma}_{\hat{\beta}_1\hat{\beta}_2} & \cdots & \hat{\sigma}_{\hat{\beta}_1\hat{\beta}_N} \\ \hat{\sigma}_{\hat{\beta}_2\hat{\beta}_1} & \hat{\sigma}^2_{\hat{\beta}_2} & \cdots & \hat{\sigma}_{\hat{\beta}_2\hat{\beta}_N} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}_{\hat{\beta}_N\hat{\beta}_1} & \hat{\sigma}_{\hat{\beta}_N\hat{\beta}_2} & \cdots & \hat{\sigma}^2_{\hat{\beta}_N} \end{pmatrix}. \quad (3)$$

Елементи цієї матриці використовуються при тестуванні гіпотез по окремим регресійним коефіцієнтам і для розрахунку інтервалу довіри.

На головній діагоналі матриці $\hat{\Sigma}_{\hat{\beta}}$ k-й елемент є 1МНК-оцінником дисперсії k-го оціненого регресійного коефіцієнта $\hat{\beta}_k$, а (k,i)-й елемент $\hat{\sigma}_{\hat{\beta}_k\hat{\beta}_i}$ є 1МНК-оцінником коваріації між $\hat{\beta}_k$ і $\hat{\beta}_i$.

Якщо оцінена коваріація $\hat{\sigma}_{\hat{\beta}_k\hat{\beta}_i} > 0$, то зв'язок між $\hat{\beta}_k$ і $\hat{\beta}_i$ прямий, тобто зі збільшенням (зменшенням) $\hat{\beta}_k$ збільшується (зменшується) $\hat{\beta}_i$, а якщо $\hat{\sigma}_{\hat{\beta}_k\hat{\beta}_i} < 0$, то зв'язок обернений – зі збільшенням (зменшенням) $\hat{\beta}_k$ зменшується (збільшується) $\hat{\beta}_i$ і навпаки.

Але коваріація має недолік, а саме – вона необмежена зверху і знизу: $-\infty < \sigma_{\beta_k \beta_i} < +\infty$. Тому до більш конкретного висновку про зв'язок між оціненими коефіцієнтами прийдемо, якщо розрахуємо коефіцієнти кореляції:

$$\hat{r}_{\hat{\beta}_k \hat{\beta}_i} = \frac{\hat{\sigma}_{\hat{\beta}_k \hat{\beta}_i}}{\hat{\sigma}_{\hat{\beta}_k} \cdot \hat{\sigma}_{\hat{\beta}_i}}, \tag{4}$$

де $\hat{\sigma}_{\hat{\beta}_k} = \sqrt{\hat{\sigma}^2_{\hat{\beta}_k}}$ – стандартне відхилення.

Приклад 5. Розрахувати оцінену дисперсійно-коваріаційну матрицю для моделі А.

Скориставшись формулою (3), отримаємо:

$$\hat{\Sigma}_{\hat{\beta}} = \hat{\sigma}^2_u \left(X^T X \right)^{-1} =$$

$$= \begin{pmatrix} \mathbf{2.873} & \mathbf{-3.617} & \mathbf{-9.473 \cdot 10^{-6}} & \mathbf{-0.081} & \mathbf{-0.047} \\ \mathbf{-3.617} & \mathbf{10.235} & \mathbf{2.127 \cdot 10^{-5}} & \mathbf{0.041} & \mathbf{-0.022} \\ \mathbf{-9.473 \cdot 10^{-6}} & \mathbf{2.127 \cdot 10^{-5}} & \mathbf{1.752 \cdot 10^{-10}} & \mathbf{-1.996 \cdot 10^{-7}} & \mathbf{-5.138 \cdot 10^{-8}} \\ \mathbf{-0.081} & \mathbf{0.041} & \mathbf{-1.996 \cdot 10^{-7}} & \mathbf{0.012} & \mathbf{-1.545 \cdot 10^{-4}} \\ \mathbf{-0.047} & \mathbf{-0.022} & \mathbf{-5.138 \cdot 10^{-8}} & \mathbf{-1.545 \cdot 10^{-4}} & \mathbf{0.003} \end{pmatrix},$$

де $\hat{\sigma}^2_{\hat{\beta}_1} = \mathbf{2.873}$, $\hat{\sigma}^2_{\hat{\beta}_2} = \mathbf{10.235}$, $\hat{\sigma}^2_{\hat{\beta}_3} = \mathbf{1.752 \cdot 10^{-10}}$, $\hat{\sigma}^2_{\hat{\beta}_4} = \mathbf{0.012}$, $\hat{\sigma}^2_{\hat{\beta}_5} = \mathbf{0.003}$,

$\hat{\sigma}_{\hat{\beta}_1 \hat{\beta}_2} = \hat{\sigma}_{\hat{\beta}_2 \hat{\beta}_1} = \mathbf{-3.617}$, $\hat{\sigma}_{\hat{\beta}_1 \hat{\beta}_3} = \hat{\sigma}_{\hat{\beta}_3 \hat{\beta}_1} = \mathbf{-9.473 \cdot 10^{-6}}$ і т. д.

Оцінені коефіцієнти кореляції розрахуємо за формулою (4). Тоді кореляційна матриця оцінених коефіцієнтів має вигляд:

$$r = \begin{pmatrix} \mathbf{1.0000} & \mathbf{-0.6669} & \mathbf{-0.4222} & \mathbf{-0.4306} & \mathbf{-0.5176} \\ \mathbf{-0.6669} & \mathbf{1.0000} & \mathbf{0.5023} & \mathbf{0.1168} & \mathbf{-0.1267} \\ \mathbf{-0.4222} & \mathbf{0.5023} & \mathbf{1.0000} & \mathbf{-0.1366} & \mathbf{-0.0724} \\ \mathbf{-0.4306} & \mathbf{0.1168} & \mathbf{-0.1366} & \mathbf{1.0000} & \mathbf{-0.0261} \\ \mathbf{-0.5176} & \mathbf{-0.1267} & \mathbf{-0.0724} & \mathbf{-0.0261} & \mathbf{1.0000} \end{pmatrix}.$$

Статистичні властивості 1МНК-оцінника $\hat{\beta}$

Оцінки регресійних коефіцієнтів, отриманих за допомогою 1МНК, – випадкові величини. Тому їх статистичні властивості важливі для оцінки методу 1МНК.

Якщо виконуються всі умови У1 – У5, то оцінник $\hat{\beta}$ має такі властивості:

1. Математичне очікування $\hat{\beta}$ дорівнює істинному значенню β, тобто $M\left(\hat{\beta}\right) = \beta$.
2. Оцінник $\hat{\beta}$ лінійний за Y: $\hat{\beta} = \left(X^T X\right)^{-1} X^T Y = A\,Y$.

3. Теорема Гауса – Маркова.

   Оцінка $\widehat{\boldsymbol{\beta}}$ – ефективна, тобто має найменшу дисперсію серед усіх можливих лінійних незміщених оцінок:

$$\widehat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\widehat{\beta}_k} \leq \widehat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\overline{\beta}_k}, \ \boldsymbol{i} = \overline{\mathbf{1}, \boldsymbol{N}},$$

   де $\widehat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\overline{\beta}_k}$ можлива оцінка $\boldsymbol{\beta_k}$ за допомогою якогось іншого методу.

4. Оцінки $\widehat{\boldsymbol{\beta}}$ переконливі, тобто при зростанні кількості спостережень $\boldsymbol{T}$ оцінки прямують за імовірністю до істинного значення:

$$\lim_{\boldsymbol{T} \to +\infty} \boldsymbol{P}\left\{\left|\widehat{\boldsymbol{\beta}}_{i\boldsymbol{T}} - \boldsymbol{\beta}_i\right| < \varepsilon\right\} = \mathbf{1}.$$

ТЕМА 4. ЗНАЧУЩІСТЬ КОЕФІЦІЄНТІВ РЕГРЕСІЇ І ЇХ ІНТЕРВАЛИ ДОВІРИ. ПРОГНОЗ РЕГРЕСАНДА. ПРОГНОЗНІ ІНТЕРВАЛИ

Розглянемо t-тест, який використовується для перевірки гіпотез про істинні, але невідомі значення окремих або декількох коефіцієнтів регресії. При цьому статистично хочемо довести, що ці коефіцієнти не дорівнюють певному значенню, що сформульовано в якості нульової гіпотези, або що істинні, але невідомі значення двох або декількох коефіцієнтів рівняння не задовольняють певному співвідношенню величин. Інтервали довіри, які досить тісно пов'язані з t-тестом, також можуть використовуватися для перевірки гіпотез про числові значення окремих коефіцієнтів регресії або їх лінійної комбінації.

Умовою проведення t-тесту є наявність таких умов:

1) вектор $U = \left(u_1, \ldots, u_T\right)^T$ є T-вимірним нормально розподіленим з нульовим вектором математичного очікування і коваріаційною матрицею $\sigma_u^2 I$;
2) регресійна матриця X детермінована і має повний ранг N.

У регресійному аналізі дуже важливим є питання: чи суттєво впливає k-й регресор у генеральній сукупності на регресанд, іншими словами, чи відрізняється істинне невідоме значення коефіцієнта регресії β_k від нуля? Для відповіді на це питання скористаємося тестом:

1. Формулюємо гіпотези: $H_A : \beta_k \neq 0,\ H_0 : \beta_k = 0$.
2. Обираємо рівень значущості $\alpha\left(0.1;\ 0.05\right)$.
3. Розраховуємо $t_{cm} = \dfrac{\widehat{\beta}_k}{\widehat{\sigma}_{\widehat{\beta}_k}}$.
4. Знаходимо $t_{kp} = t\left(\dfrac{\alpha}{2}, T - N\right)$ за таблицею Ст'юдента.
5. Якщо $\left|t_{cm}\right| > t_{kp}$, то гіпотеза H_0 відхиляється з ймовірністю $\left(1 - \alpha\right)100\%$ і приймається гіпотеза H_A.

Розглянемо більш загальний тест з гіпотезами: $H_A : \widehat{\beta}_k \neq \beta_k^0,\ H_0 : \widehat{\beta}_k = \beta_k^0$, де β_k^0 є довільне, наперед встановлене і логічно обґрунтоване реальне число (випадок, коли $\beta_k^0 = 0$, розглянуто вище). t_{cm} розраховується таким чином:

$$t_{cm} = \frac{\widehat{\beta}_k - \beta_k^0}{\widehat{\sigma}_{\widehat{\beta}_k}}.$$

Якщо $\left|t_{cm}\right| > t_{kp} = t\left(\frac{\alpha}{2}, T-N\right)$, то з ймовірністю $(1-\alpha)100\%$ гіпотеза H_0 відхиляється і приймається гіпотеза H_A.

t-тест гіпотез про лінійну комбінацію коефіцієнтів регресії

У регресійному аналізі можуть виникнути такі питання:

1. Чи є рівновеликими два регресійні коефіцієнти?
2. Чи є сума декількох регресійних коефіцієнтів постійною величиною?

Для відповіді на ці питання можна скористатися тестом:

1. Формулюються гіпотези:

$$H_A : c_1\beta_1 + c_2\beta_2 + \ldots + c_N\beta_N \neq c^0, \text{ або } c^T\beta \neq c^0;$$

$$H_0 : c_1\beta_1 + c_2\beta_2 + \ldots + c_N\beta_N = c^0, \text{ або } c^T\beta = c^0,$$

де $c_i\left(i = \overline{1,N}\right)$ – константи, значення яких задаються,

β_k – істинні невідомі значення регресійних коефіцієнтів,

c^0 – скаляр, значення якого задається.

2. Задається рівень значущості α.
3. Розраховується $t_{cm} = \frac{c^T\widehat{\beta} - c^0}{\widehat{\sigma}}$, де $\widehat{\sigma} = \sqrt{c^T \cdot \widehat{\Sigma}_{\widehat{\beta}} \cdot c}$.
4. Визначаємо $t_{kp} = t\left(\frac{\alpha}{2}, T-N\right)$.
5. Якщо $\left|t_{cm}\right| > t_{kp}$, то з ймовірністю $(1-\alpha)100\%$ гіпотеза H_0 відхиляється.

Отримані оцінки коефіцієнтів регресії називають точковими, так як на числовій осі значення цього коефіцієнта представляє відповідна точка. Може виникнути питання: а наскільки ця точкова оцінка відрізняється від відповідного істинного значення? Для відповіді на це питання необхідно побудувати інтервал довіри для коефіцієнта.

Інтервал довіри (інтервальна оцінка) для регресійного коефіцієнта β_k при рівні довіри $(1-\alpha)$ є інтервалом з випадковими межами і включає істинне значення k-го регресійного коефіцієнта з ймовірністю $(1-\alpha)100\%$.

Він має такі межі:

$$\text{Н. М.: } \widehat{\beta}_k - t\left(\frac{\alpha}{2}, T-N\right)\widehat{\sigma}_{\widehat{\beta}_k},$$

$$\text{В. М.: } \widehat{\beta}_k + t\left(\frac{\alpha}{2}, T-N\right)\widehat{\sigma}_{\widehat{\beta}_k}.$$

Знаючи інтервали довіри для значень k-го коефіцієнта регресії, можна відповісти на питання про значущість цього коефіцієнта, а саме: якщо нуль належить інтервалу довіри, то коефіцієнт незначимо відрізняється від нуля з ймовірністю $(1-\alpha)100\%$.

Якщо деякі коефіцієнти моделі незначимо відрізняються від нуля, то цю модель не можна використовувати для прогнозу, а тільки для досліджень. Не варто користуватися моделлю, якщо всі коефіцієнти незначимо відрізняються від нуля.

Приклад 6. Визначити, чи значимо відрізняються від нуля коефіцієнти моделі А і розрахувати інтервали довіри для коефіцієнтів цієї моделі. Візьмемо $\alpha = 0.1$.

Розглядається модель А:

$$\hat{y} = 10.439\,x_1 - 14.755\,x_2 + 0.00003\,x_3 + 0.198\,x_4 - 0.00009\,x_5.$$

Розрахуємо $t_{kp} = t\left(\frac{\alpha}{2},\, T-N\right) = t\left(\frac{0.1}{2},\, 25-5\right) = 1.725$.

Далі розраховуємо t_{cm}:

$$\beta_1 : t_{cm} = \frac{\widehat{\beta}_1}{\widehat{\sigma}_{\widehat{\beta}_1}} = \frac{10.439}{\sqrt{2.873}} = 6.159,$$

$$\beta_2 : t_{cm} = \frac{\widehat{\beta}_2}{\widehat{\sigma}_{\widehat{\beta}_2}} = \frac{-14.755}{\sqrt{10.235}} = -4.612,$$

$$\beta_3 : t_{cm} = \frac{\widehat{\beta}_3}{\widehat{\sigma}_{\widehat{\beta}_3}} = \frac{0.00003}{\sqrt{1.752 \cdot 10^{-10}}} = 2.364,$$

$$\beta_4 : t_{cm} = \frac{\widehat{\beta}_4}{\widehat{\sigma}_{\widehat{\beta}_4}} = \frac{0.198}{\sqrt{0.012}} = 1.796,$$

$$\beta_5 : t_{cm} = \frac{\widehat{\beta}_5}{\widehat{\sigma}_{\widehat{\beta}_5}} = \frac{-0.00009}{\sqrt{0.003}} = -0.002.$$

Так як $\left|t_{cm}\right|$ для $\beta_1, \beta_2, \beta_3, \beta_4$ більші за t_{kp}, то $\beta_1, \beta_2, \beta_3, \beta_4$ значимо відрізняються від нуля з ймовірністю 90%. У β_5 $\left|t_{cm}\right|$ менше t_{kp} і тому коефіцієнт β_5 незначимо відрізняється від нуля, тобто регресор x_5 не впливає на регресанд.

Визначимо інтервали довіри для коефіцієнтів моделі А:

$$\beta_1 : H.M. : \widehat{\beta}_1 - t\left(\frac{\alpha}{2},\, T-N\right)\widehat{\sigma}_{\widehat{\beta}_1} = 10.439 - 1.725\sqrt{2.873} = 7.515,$$

$$B.M. : \widehat{\beta}_1 + t\left(\frac{\alpha}{2},\, T-N\right)\widehat{\sigma}_{\widehat{\beta}_1} = 10.439 + 1.725\sqrt{2.873} = 13.363,$$

$$\beta_2 : H.M.: \widehat{\beta}_2 - t\left(\frac{\alpha}{2}, T-N\right)\widehat{\sigma}_{\widehat{\beta}_2} = -14.755 - 1.725\sqrt{10.235} = -20.273,$$

$$B.M.: \widehat{\beta}_2 + t\left(\frac{\alpha}{2}, T-N\right)\widehat{\sigma}_{\widehat{\beta}_2} = -14.755 + 1.725\sqrt{10.235} = -9.236,$$

$$\beta_3 : H.M.: \widehat{\beta}_3 - t\left(\frac{\alpha}{2}, T-N\right)\widehat{\sigma}_{\widehat{\beta}_3} = 0.00003 - 1.725\sqrt{1.752 \cdot 10^{-10}} =$$

$$= 0.000008,$$

$$B.M.: \widehat{\beta}_3 + t\left(\frac{\alpha}{2}, T-N\right)\widehat{\sigma}_{\widehat{\beta}_3} = 0.00003 + 1.725\sqrt{1.752 \cdot 10^{-10}} =$$

$$= 0.00005,$$

$$\beta_4 : H.M.: \widehat{\beta}_4 - t\left(\frac{\alpha}{2}, T-N\right)\widehat{\sigma}_{\widehat{\beta}_4} = 0.198 - 1.725\sqrt{0.012} = 0.008,$$

$$B.M.: \widehat{\beta}_4 + t\left(\frac{\alpha}{2}, T-N\right)\widehat{\sigma}_{\widehat{\beta}_4} = 0.198 + 1.725\sqrt{0.012} = 0.389,$$

$$\beta_5 : H.M.: \widehat{\beta}_5 - t\left(\frac{\alpha}{2}, T-N\right)\widehat{\sigma}_{\widehat{\beta}_5} = -0.00009 - 1.725\sqrt{0.003} = -0.093,$$

$$B.M.: \widehat{\beta}_5 + t\left(\frac{\alpha}{2}, T-N\right)\widehat{\sigma}_{\widehat{\beta}_5} = -0.00009 + 1.725\sqrt{0.003} = 0.092.$$

Інтерпретація для β_1: істинне значення β_1 з ймовірністю 90% буде коливатися від 7.515 до 13.363.

Так як інтервалам довіри для $\beta_1, \beta_2, \beta_3, \beta_4$ нуль не належить, то ці коефіцієнти значимо відрізняються від нуля з ймовірністю 80%.

Якщо розглянути інтервал довіри для β_5, то бачимо, що нуль належить йому і тому β_5 незначимо відрізняється від нуля з ймовірністю 90%, що ми раніше уже отримали.

Точкові і інтервальні прогнози регресанда

Прогноз робиться для значень регресорів, що взяті з тієї ж генеральної сукупності, що і значення їх для оцінки β_k.

Якість прогнозу залежить від надійності оцінок коефіцієнтів моделі, повноти виконання передумов 1МНК і якості оцінок значень регресорів для прогнозної точки.

Прогноз при $x = \tilde{x}$ розраховується за формулою

$$\hat{y}^{\Pi} = \hat{\beta}^T \tilde{x}. \qquad (1)$$

Точкове оцінене значення $\hat{y}^{\Pi}$, що розраховане за допомогою (1), може мати такі дві інтерпретації:

а) оцінку математичного очікування регресанда;

б) оцінку індивідуального значення регресанда.

Зазначимо, що в економічних дослідженнях важливе значення має не точкова оцінка прогнозу регресанда, а прогнозний інтервал.

Прогнозний інтервал величини математичного очікування регресанда має такі межі:

$$H.M.: \hat{y}^{\Pi} - t\left(\alpha, T-N\right)\hat{\sigma}_e,$$

$$B.M.: \hat{y}^{\Pi} + t\left(\alpha, T-N\right)\hat{\sigma}_e,$$

де $\hat{y}^{\Pi}$ розраховується за формулою (1), а $\hat{\sigma}_e = \sqrt{\tilde{x}^T \cdot \hat{\Sigma}_{\hat{\beta}} \cdot \tilde{x}}$, $\tilde{x}$ – прогнозна точка.

Інтервал прогнозу (прогнозний інтервал) для індивідуального значення регресанда розраховується так:

$$H.M.: \hat{y}^{\Pi} - t\left(\alpha, T-N\right)\hat{\sigma}_{e(i)},$$

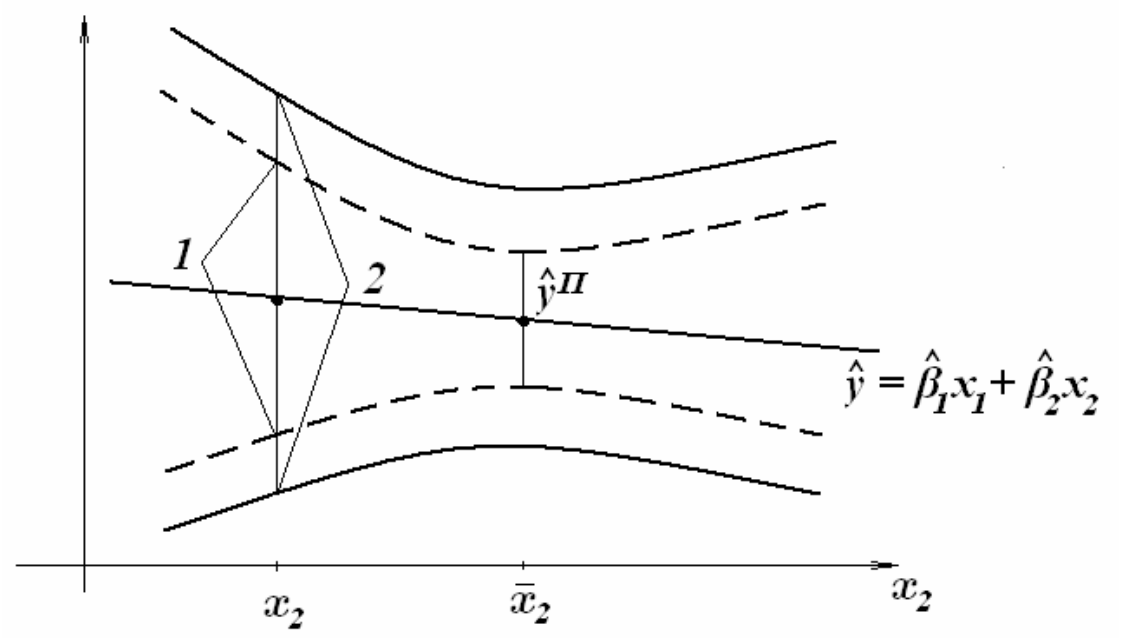
$$B.M.: \hat{y}^{\Pi} + t\left(\alpha, T-N\right)\hat{\sigma}_{e(i)},$$

де $\hat{\sigma}_{e(i)} = \sqrt{\hat{\sigma}_u^2 + \tilde{x}^T \cdot \hat{\Sigma}_{\hat{\beta}} \cdot \tilde{x}}$.

Так як $\hat{\sigma}_u^2 \geq 0$, то прогнозний інтервал індивідуального значення завжди містить в собі прогнозний інтервал математичного очікування регресанда.

Зазначимо, що обидва інтервали будуть найменшими при $\tilde{x} = \bar{x}$ ($\bar{x}$ – вектор середніх значень x).

Розглянемо важливий зв'язок між точковим прогнозом і обома інтервальними прогнозами для гіпотетичної парної регресії в залежності від значення регресора x_2 в прогнозному періоді.



Позначення:

———— межі прогнозного інтервалу індивідуального значення регресанда,

---------- межі прогнозного інтервалу математичного очікування регресанда,

1 величина прогнозного інтервалу математичного очікування регресанда,

2 величина прогнозного інтервалу індивідуального значення регресанда.

Як обирати прогнозні значення $\tilde{x}_i$? Якщо відомо, що $a_i \le \tilde{x}_i \le b_i$, $i = \overline{2, N}$, то прогнозні значення $\tilde{x}_i$ обирають так, щоб

$$a_i - \frac{b_i - a_i}{3} \le \tilde{x}_i \le b_i + \frac{b_i - a_i}{3}.$$

Приклад 7. Розрахувати точковий і інтервальний прогнози для моделі А при $\tilde{x} = \bar{x}$ ($\bar{x}$ – вектор середніх значень x).

Нагадаємо, що $\bar{x} = \left(1;\ 0.308;\ 29587.28;\ 6.301;\ 19.550\right)^T$. Тоді

$$\hat{y}^{\text{П}} = 10.439 \cdot 1 - 14.755 \cdot 0.308 + 0.00003 \cdot 29587.28 + 0.198 \cdot 6.301 - \\ - 0.00009 \cdot 19.550 = 8.068.$$

Прогнозне значення $\hat{y}^{\text{П}} = 8.068$ має такі інтерпретації:

1) середня продуктивність праці на підприємствах, у яких трудомісткість продукції дорівнює 0.308, середньорічний фонд заробітної платні підприємства – 29587.28, фондоозброєність праці – 6.301, невиробничі витрати – 19.550, дорівнює 8.068;
2) продуктивність праці на підприємстві, у якого трудомісткість продукції дорівнює 0.308, середньорічний фонд заробітної платні підприємства – 29587.28, фондоозброєність праці – 6.301, невиробничі витрати – 19.550, дорівнює 8.068.

Оцінимо інтервал довіри для математичного очікування. Маємо

$$\hat{\sigma}_e = \sqrt{\tilde{x}^T \cdot \hat{\Sigma}_{\hat{\beta}} \cdot \tilde{x}} =$$

$$= \sqrt{\begin{pmatrix} 1 \\ 0.308 \\ 29587.28 \\ 6.301 \\ 19.550 \end{pmatrix}^T \begin{pmatrix} 2.873 & -3.617 & -9.473 \cdot 10^{-6} & -0.081 & -0.047 \\ -3.617 & 10.235 & 2.127 \cdot 10^{-5} & 0.041 & -0.022 \\ -9.473 \cdot 10^{-6} & 2.127 \cdot 10^{-5} & 1.752 \cdot 10^{-10} & -1.996 \cdot 10^{-7} & -5.138 \cdot 10^{-8} \\ -0.081 & 0.041 & -1.996 \cdot 10^{-7} & 0.012 & -1.545 \cdot 10^{-4} \\ -0.047 & -0.022 & -5.138 \cdot 10^{-8} & -1.545 \cdot 10^{-4} & 0.003 \end{pmatrix} \begin{pmatrix} 1 \\ 0.308 \\ 29587.28 \\ 6.301 \\ 19.550 \end{pmatrix}} =$$

$$= 0.226.$$

Тоді

$$H.M.: \hat{y}^{\text{П}} - t\left(\alpha, T - N\right)\hat{\sigma}_e = 8.068 - 1.725 \cdot 0.226 = 7.678,$$

$$B.M.: \hat{y}^{\Pi} + t\Big(\alpha, T-N\Big)\widehat{\sigma}_e = 8.068 + 1.725 \cdot 0.226 = 8.458.$$

Отже, середня продуктивність праці на підприємствах, у яких трудомісткість продукції дорівнює 0.308, середньорічний фонд заробітної платні підприємства – 29587.28, фондоозброєність праці – 6.301, невиробничі витрати – 19.550, буде коливатися з ймовірністю 80% у межах від 7.678 до 8.458.

Тепер оцінимо прогнозний інтервал для індивідуального значення регресанда. Розрахуємо

$$\widehat{\sigma}_{e(i)} = \sqrt{\widehat{\sigma}_u^2 + \tilde{x}^T \cdot \widehat{\Sigma}_{\widehat{\beta}} \cdot \tilde{x}} = \sqrt{1.280 + 0.226} = 1.154.$$

Тоді

$$H.M.: \hat{y}^{\Pi} - t\Big(\alpha, T-N\Big)\widehat{\sigma}_{e(i)} = 8.068 - 1.725 \cdot 1.154 = 6.078,$$

$$B.M.: \hat{y}^{\Pi} + t\Big(\alpha, T-N\Big)\widehat{\sigma}_{e(i)} = 8.068 + 1.725 \cdot 1.154 = 10.058.$$

Отримали, що продуктивність праці на підприємстві, у якого трудомісткість продукції дорівнює 0.308, середньорічний фонд заробітної платні підприємства – 29587.28, фондоозброєність праці – 6.301, невиробничі витрати – 19.550, з ймовірністю 80% буде коливатися від 6.078 до 10.058.

Як і очікували, прогнозний інтервал для математичного очікування вужчий за прогнозний інтервал для індивідуального значення.

ТЕМА 5. ПОКАЗНИКИ АДЕКВАТНОСТІ РЕГРЕСІЙНОЇ МОДЕЛІ

Коефіцієнт детермінації

У класичному регресійному аналізі вважається, що функція регресії відома до оцінки значень параметрів, тобто лінійна регресійна модель правильно специфікована. Однак в емпіричних економічних і соціальних дослідженнях при застосуванні регресійного аналізу перш за все повинна бути обрана з множини варіантів рівнянь (які відрізняються регресорами) найбільш адекватна регресійна модель.

Для оцінки коефіцієнтів регресії використовується сума квадратів залишків. Її ми використовували для порівняння декількох моделей. Однак її не можна використовувати як показник адекватності однієї моделі, так як вона необмежена зверху. Відсутність верхньої межи у сумі квадратів залишків, як недолік, усувається за допомогою коефіцієнта детермінації.

Коефіцієнт детермінації R^2 можна ввести як квадрат емпіричного коефіцієнта кореляції між двома рядами спостережень: емпіричних даних y_t і розрахованих $\hat{y}_t$. Тобто

$$R^2 = r^2 = \frac{\left[\sum_{t=1}^{T}\left(y_t - \bar{y}\right)\left(\hat{y}_t - \bar{y}\right)\right]^2}{\sum_{t=1}^{T}\left(y_t - \bar{y}\right)^2 \sum_{t=1}^{T}\left(\hat{y}_t - \bar{y}\right)^2}. \quad (1)$$

Зазначимо, що $\bar{y} = \bar{\hat{y}}$. Друге рівноцінне визначення коефіцієнта детермінації робиться таким чином:

$$R^2 = \frac{\sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2}. \quad (2)$$

І нарешті третє визначення:

$$R^2 = 1 - \frac{\sum_{t=1}^{T} u_t^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2}. \quad (3)$$

Коефіцієнт детермінації належить проміжку $\left[0;\ 1\right]$.

Чим ближче R^2 до одиниці, тим краще регресійна модель апроксимує емпіричні дані. Якщо $\sum_{t=1}^{T} u_t^{\,2} = 0$, то $R^2 = 1$. У цьому випадку $y_t = \hat{y}_t$, $t = \overline{1,T}$, а це означає, що всі

емпіричні дані розташовані на регресійній гіперплощині. Випадок $R^2 = 0$ можливий при $\hat{y}_t = \bar{y}$ для $t = \overline{1,T}$.

З двох регресійних рівнянь, які відрізняються лише регресорами (кількість їх однакова), кращим вважається рівняння з найбільшим коефіцієнтом детермінації.

Коефіцієнт детермінації можна використати для визначення значущості регресійного рівняння, для цього розраховуємо $F_{cm} = \dfrac{R^2(T-N)}{(1-R^2)(N-1)}$ і якщо $F_{cm} > F_{kp} = F(\alpha, N-1, T-N)$, то рівняння значимо відрізняється від нуля з імовірністю $(1-\alpha)100\%$.

Скоригований за Тейлом коефіцієнт детермінації

Звичайний коефіцієнт детермінації R^2 як критерій вибору функції регресії має недолік, який може призвести до того, що буде віддаватися перевага варіанту рівняння з великою кількістю регресорів. Цей недолік полягає в тому, що при включенні додаткового регресору до рівняння R^2 може тільки збільшитись. Якщо якість різних варіантів регресійного рівняння оцінювати тільки за допомогою R^2, то рівняння з більшою кількістю регресорів, як правило, буде давати кращі результати, ніж з відносно малою їх кількістю. Однак з кожним додатковим регресором губиться одна ступінь вільностей, а цей недолік, на жаль, не враховується, коли R^2 виступає у якості критерію вибору. Кількість ступенів свободи для регресійної моделі визначається як $T - N$, де T – кількість спостережень, а N – кількість регресорів. Якщо до регресійного рівняння включити додатковий регресор, то кількість регресорів N у рівнянні збільшується на 1, а кількість ступенів вільності зменшується на 1. При застосуванні t- і F-тестів, а також при побудові інтервалів довіри і прогнозних, бажано мати по можливості більшу кількість ступенів свободи. Чим більша кількість ступенів свободи, тим менші будуть інтервали. Тому у статистичному відношенні наявність додаткового регресора не завжди може бути бажана.

Що ж робити, коли ми хочемо порівнювати моделі з різною кількістю регресорів? У таких випадках необхідно коригувати коефіцієнт детермінації з урахуванням кількості регресорів x, які входять до різних моделей, тобто зменшити вплив залежності значення коефіцієнта детермінації від кількості регресорів. Для цього вводиться спеціальний скоригований за Тейлом коефіцієнт детермінації, який має вигляд:

$$\bar{R}^2_T = 1 - \frac{\dfrac{\sum_{t=1}^{T} u_t^2}{T-N}}{\dfrac{\sum_{t=1}^{T}(y_t - \bar{y})^2}{T-1}}. \qquad (4)$$

На відміну від простого коефіцієнта детермінації R^2, скоригований за Тейлом коефіцієнт детермінації коригується з урахуванням ступенів вільностей суми квадратів залишків та загальної суми квадратів.

Скоригований коефіцієнт детермінації за Тейлом пов'язаний з коефіцієнтом детермінації R^2 таким чином:

$$\overline{R}_T^2 = 1 - \left(1 - R^2\right)\frac{T-1}{T-N}. \quad (5)$$

Зазначимо, що порівнювати значення двох або більше коефіцієнтів детермінації R^2 чи $\overline{R}_T^2$ можна лише за однакових залежних змінних, які можуть набирати різних функціональних форм.

Розглянемо деякі властивості скоригованого коефіцієнта детермінації за Тейлом $\overline{R}_T^2$:

1) при $N > 1$ скоригований коефіцієнт детермінації за Тейлом не більший за звичайний коефіцієнт детермінації, тобто $\overline{R}_T^2 \leq R^2$;
2) якщо кількість регресорів зростає, то скоригований коефіцієнт детермінації за Тейлом зростає повільніше, ніж R^2, тобто зменшується вплив кількості чинників на величину коефіцієнта детермінації;
3) скоригований за Тейлом коефіцієнт детермінації $\overline{R}_T^2$ може бути і від'ємним, тоді як R^2 завжди невід'ємний;
4) якщо до регресії додається новий чинник і його $\left|t_{cm}\right| > 1$, то скоригований за Тейлом коефіцієнт детермінації $\overline{R}_T^2$ збільшується, і тільки у цьому випадку.

Збільшення скоригованого за Тейлом коефіцієнта детермінації $\overline{R}_T^2$ при введенні нового регресора до моделі не обов'язково означає, що його коефіцієнт значимо відрізняється від нуля. Тому не потрібно вважати, що збільшення $\overline{R}_T^2$ означає покращення специфікації моделі. Це і є однією з причин, що заважає широкому використанню скоригованого за Тейлом коефіцієнта детермінації.

В останній час увага до коефіцієнта детермінації R^2 зменшилася, тому що у погано специфікованої моделі може бути великим коефіцієнт детермінації (наприклад, при наявності мультиколінеарності). Тепер дослідники розглядають коефіцієнт детермінації R^2 як один з діагностичних показників, що розраховуються при побудові моделі.

Скоригований за Амемією коефіцієнт детермінації

Скоригований коефіцієнт детермінації за Амемією $\overline{R}_A^2$ розраховується таким чином:

$$\overline{R}_A^2 = 1 - \left(1 - R^2\right)\frac{T+N}{T-N}. \quad (6)$$

Скоригований коефіцієнт детермінації $\overline{R}^2_A$ відображає втрату ступеня свободи при включенні додаткового регресора більш чітко, ніж $\overline{R}^2_T$. Це означає, що $\overline{R}^2_A$ змінюється на більшу величину, ніж $\overline{R}^2_T$, при включенні додаткового регресора. Тому той, хто застосовує $\overline{R}^2_A$ замість $\overline{R}^2_T$ у якості критерію вибору, буде, при інших рівних умовах, віддавати перевагу рівнянню, що має меншу кількість регресорів.

Частковий коефіцієнт детермінації

Відомо, що $R^2(k+1) \geq R^2(k)$, тобто якщо включити до моделі регресор, то коефіцієнт детермінації не може зменшитися. Це означає, що

$$\Delta R^2 \text{ (від додаткового регресора) =}$$

$$= R^2 \text{ (з } k+1 \text{ регресором)} - R^2 \text{ (з } k \text{ регресорами).}$$

Щоб кількісно визначити ΔR^2, необхідно оцінити дві регресії: одну з k, а другу з $(k+1)$ регресорами. А можна піти іншим шляхом: без значних обчислювальних витрат розрахувати коефіцієнт ΔR^2_k, який називається частковим коефіцієнтом детермінації:

$$\Delta R^2_k = \Delta R^2_{x_k} = \frac{1-R^2}{T-N}\left(\frac{\widehat{\beta}_k}{\widehat{\sigma}_{\widehat{\beta}_k}}\right)^2. \qquad (7)$$

Інтерпретація ΔR^2_k: він показує, на яку величину зменшиться коефіцієнт детермінації, якщо k-й регресор виключити з моделі.

Враховуючи суть коефіцієнта детермінації, можна зробити такий висновок: чим більший ΔR^2_k, тим більш важливим є у моделі k-й регресор.

Отже, ми отримали ще одну ознаку впливовості регресора в моделі.

Приклад 8. Для моделей А, Б, В, Г, Д розрахувати коефіцієнти детермінації R^2, $\overline{R}^2_T$, $\overline{R}^2_A$, а для моделі А – часткові коефіцієнти детермінації ΔR^2_2, ΔR^2_3, ΔR^2_4, ΔR^2_5. Визначити, які моделі значимо відрізняються від нуля, а які ні.

Розрахуємо всі параметри для моделі А.

Маємо:

$$R^2 = \frac{\sum_{t=1}^{25}(\hat{y}_t - \bar{y})^2}{\sum_{t=1}^{25}(y_t - \bar{y})^2} = 0.7628.$$

Коефіцієнт детермінації досить великий. Розрахуємо $\overline{R}^2_T$:

$$\overline{R}_T^2 = 1 - \left(1 - R^2\right)\frac{T-1}{T-N} = 1 - \left(1 - 0.7628\right)\frac{25-1}{25-5} = 0.7154.$$

Тепер розрахуємо $\overline{R}_A^2$:

$$\overline{R}_A^2 = 1 - \left(1 - R^2\right)\frac{T+N}{T-N} = 1 - \left(1 - 0.7628\right)\frac{25+5}{25-5} = 0.6442.$$

Як і писали вище, виконується співвідношення

$$R^2 > \overline{R}_T^2 > \overline{R}_A^2.$$

Визначимо, чи значимо відрізняється від нуля модель А. Для цього розрахуємо F_{cm} і F_{kp}:

$$F_{cm} = \frac{R^2(T-N)}{(1-R^2)(N-1)} = \frac{0.7628\left(25-5\right)}{\left(1-0.7628\right)\left(5-1\right)} = 16.082,$$

$$F_{kp} = F(\alpha, N-1, T-N) = F\left(0.05,\ 4,\ 20\right) = 2.87.$$

Так як $F_{cm} > F_{kp}$, то з ймовірністю 95% модель А значимо відрізняється від нуля.

Такі ж параметри для інших моделей розраховуються аналогічно.

Тепер розрахуємо часткові коефіцієнти детермінації для моделі А.

Отримаємо:

$$\Delta R_2^2 = \frac{1-R^2}{T-N}\left(\frac{\widehat{\beta}_2}{\widehat{\sigma}_{\widehat{\beta}_2}}\right)^2 = \frac{1-0.7628}{25-5}\left(\frac{-14.755}{\sqrt{10.235}}\right)^2 = 0.252,$$

$$\Delta R_3^2 = \frac{1-R^2}{T-N}\left(\frac{\widehat{\beta}_3}{\widehat{\sigma}_{\widehat{\beta}_3}}\right)^2 = \frac{1-0.7628}{25-5}\left(\frac{0.00003}{\sqrt{1.752\cdot 10^{-10}}}\right)^2 = 0.066,$$

$$\Delta R_4^2 = \frac{1-R^2}{T-N}\left(\frac{\widehat{\beta}_4}{\widehat{\sigma}_{\widehat{\beta}_4}}\right)^2 = \frac{1-0.7628}{25-5}\left(\frac{0.198}{\sqrt{0.012}}\right)^2 = 0.038,$$

$$\Delta R_5^2 = \frac{1-R^2}{T-N}\left(\frac{\widehat{\beta}_5}{\widehat{\sigma}_{\widehat{\beta}_5}}\right)^2 = \frac{1-0.7628}{25-5}\left(\frac{-0.00009}{\sqrt{0.003}}\right)^2 = 0.00000003.$$

Зважаючи на величини ΔR_i^2, можемо зробити висновок, що найбільш впливовим регресором є x_2, потім x_3, x_4 і x_5, що співпадає з раніше отриманими результатами.

Отримані результати зведемо до однієї таблиці.

Модель		R^2	$\overline{R}_T^2$	$\overline{R}_A^2$	F_{cm}	F_{kp}
А	$x_1 - x_5$	0.7628	0.7154	0.6442	16.082	2.87
Б	x_1, x_2	0.6416	0.6260	0.4624	41.173	4.28
В	x_1, x_3	0.4413	0.4170	0.362	18.166	4.28
Г	x_1, x_4	0.1598	0.1233	-0.2603	4.374	4.28
Д	x_1, x_5	0.0028	-0.0406	-0.4958	0.064	4.28

Усі моделі, крім моделі Д, значимо відрізняються від нуля.

ТЕМА 6. СПЕЦИФІКАЦІЯ МОДЕЛІ. ПОБУДОВА НАЙКРАЩОЇ МОДЕЛІ

Побудова економетричної моделі починається з питання про специфікацію моделі. Проблема специфікації моделі включає такі питання:

1) визначення кінцевих цілей моделювання (прогноз, управління, імітація різних сценаріїв розвитку системи);
2) визначення списку екзогенних і ендогенних змінних;
3) вибір виду рівняння регресії.

Відносно першого пункту. У залежності від кінцевої мети моделювання по-різному відносяться до кількості чинників у моделі, до виконання тих чи інших передумов використання 1МНК. Наприклад, якщо хочемо зробити за моделлю прогноз, то необхідно побільше включити до моделі чинників, а якщо проводити дослідження, то навпаки. Прогноз інколи можна робити і при наявності мультиколінеарності чинників, а от дослідження моделі робити недоцільно.

Отже, необхідно чітко сформулювати для себе мету побудови економетричної моделі і переходити до вибору змінних моделі.

Змінні, що включаються до моделі, повинні задовольняти таким вимогам:

1) їх можна кількісно вимірювати, а якщо до моделі включається якісний чинник, то необхідно надати йому кількісної визначеності;
2) незалежні змінні не повинні корелювати між собою.

Оптимальний вибір кількості чинників має велике значення при побудові моделі. Відбір чинників роблять за допомогою теоретико-економічного аналізу. Якщо це неможливо зробити, то спочатку відбір роблять, виходячи із сутті проблеми, а далі на підставі коефіцієнтів кореляції між чинниками і результуючим показником. Тобто до моделі включають чинники, що мають найбільші за модулем часткові коефіцієнти кореляції з результуючим показником. При цьому слідкуємо, щоб кореляція між чинниками не була суттєвою. Можна скористатися методом Фаррара –Глобера для невключення до моделі чинників, що викликають мультиколінеарність.

При побудові моделі слід пам’ятати, що насичення моделі «зайвими» чинниками не тільки не збільшує суттєво коефіцієнт детермінації, але і призводить до зменшення точності оцінок. Елементи коваріаційної матриці для $\widehat{\boldsymbol{\beta}}$ будуть завищеними, що зменшує $\boldsymbol{t}$-статистики і коефіцієнт детермінації. Збільшуються інтервали довіри параметрів і прогнозний інтервал регресанда. Якщо з моделі видалити регресор і при цьому скоригований за Тейлом коефіцієнт детермінації збільшується, то цей регресор є зайвим.

Існують ще так звані «пропущені» змінні. Це ті суттєві змінні, що помилково не були включені до економетричної моделі. Якщо у регресійному рівнянні пропущена змінна, то оцінки коефіцієнтів регресії будуть зміщеними.

Ознакою, що вказує на пропущену екзогенну змінну, є додатні знаки у добутку оцінки коефіцієнта при пропущеній змінній з коефіцієнтом парної кореляції пропущеної змінної з усіма екзогенними змінними, включеними до моделі.

Наприклад, якщо у моделі

$$\boldsymbol{y = \beta_1 x_1 + \ldots + \beta_{m-1} x_{m-1} + \beta_{m+1} x_{m+1} + \ldots + \beta_N x_N + u}$$

пропущена змінна $\boldsymbol{x_m}$, то роблять висновок про її необхідність, якщо

$$\left\{ \beta_m \cdot r_{x_i x_m} \right\} > 0, \quad i = \overline{2, N},\ i \neq m.$$

При дослідженні специфікації моделі необхідно дотримуватися таких правил:

1) за допомогою часткових коефіцієнтів кореляції вияснити, які незалежні змінні суттєво корелюють з залежною;
2) перевірити значущість коефіцієнтів моделі;
3) вияснити, чи збільшується скоригований за Тейлом коефіцієнт детермінації при включенні до моделі нової змінної;
4) чи суттєво впливає нова змінна на оцінки коефіцієнтів при інших змінних.

Існують статистичні критерії специфікації, розглянемо два найбільш відомі.

Критерій Рамсея. Алгоритм його такий:

1) оцінюємо коефіцієнти моделі за допомогою 1МНК

$$\hat{y} = \sum_{i=1}^{N} \hat{\beta}_i x_i; \tag{1}$$

2) розраховуємо $\hat{y}_t^2, \hat{y}_t^3, \hat{y}_t^4, t = \overline{1, T}$ і оцінюємо коефіцієнти моделі

$$y = \sum_{i=1}^{N} \alpha_i x_i + \alpha_{N+1} \hat{y}^2 + \alpha_{N+2} \hat{y}^3 + \alpha_{N+3} \hat{y}^4 + \varepsilon \tag{2}$$

методом 1МНК;

3) визначаємо F_{cm}:

$$F_{cm} = \frac{(RSS_1 - RSS_2)(T-3)}{3 \cdot RSS_2},$$

де $RSS_i = \sum_{t=1}^{T} u_{it}^2,\ i = 1, 2$, $u_{i\bullet}$ – залишки в моделі (1) або (2);

4) якщо $F_{cm} > F_{kp} = F(\alpha; 3; T-3)$, то модель (1) є погано специфікованою.

Критерій Амемії. Будуємо для конкретної моделі функцію

$$AF = \frac{\sum_{i=1}^{T} u_i^2 (T+N)}{T-N}.$$

Модель, для якої значення AF є меншим, вважається краще специфікованою.

Побудова «найкращої» регресії

Далі мова буде йти про побудову лінійної регресійної моделі. Існує дві протилежні думки для вибору регресії:

1) якщо за регресійним рівнянням необхідно робити прогноз, то до рівняння включається побільше незалежних змінних, що дає більш надійний прогноз;
2) оскільки збільшення кількості незалежних змінних призводить до небажаних наслідків (див. вище) і витрати на отримання інформації збільшуються, то намагаються включити до моделі як можна менше змінних.

Компромісним рішенням якраз і є побудова «найкращого» рівняння регресії. Існує декілька способів побудови такої регресії. Розглянемо один із них – метод усіх можливих регресій. Цей метод використовує три критерії оцінки регресії: скоригований за Тейлом коефіцієнт детермінації R_T^2, залишковий середній квадрат $\hat{\sigma}_u^2$ і критерій Маллоуза – C_p-статистику.

Ідея методу дуже проста. Усі можливі регресії, що будуються на основі N регресорів, а їх буде 2^{N-1}, розбиваємо на групи. До кожної групи долучаються регресії з однаковою кількістю регресорів. За допомогою кожного з трьох критеріїв у групі знаходимо «найкращу» модель. Порівняльний аналіз результатів застосованих критеріїв дає можливість отримати «найкращу» регресію.

Зазначимо, що C_p-статистика Маллоуза розраховується за формулою

$$C_p = \frac{RSS_p}{\hat{\sigma}_u^2} - (T - 2N),$$

де p – кількість чинників у регресії, $\hat{\sigma}_u^2$ розраховано для найдовшої регресії.

Далі ідею методу проілюструємо на конкретному прикладі.

Приклад 9. Будуємо «найкращу модель» методом усіх регресій. Всього регресій буде 16. До групи А віднесемо регресію $y = f(x_1) = \bar{y}$. Зрозуміло, що ця модель не може скласти конкуренцію іншим і тому її надалі розглядати не будемо. До групи Б включаємо двофакторні регресії, В – трифакторні, Г – чотирифакторні і група Д складатиметься з однієї найдовшої регресії: $y = f(x_1, x_2, x_3, x_4, x_5)$. Спочатку порівнюємо регресії за допомогою R^2-критерію. Результати зведемо до таблиці.

	Вид регресії				R_T^2			
Б	12	13	14	15	0.6260	0.4170	0.1233	-0.0406
В	123	124	125		0.6995	0.6686	0.6098	
	134	135	145		0.4616	0.3946	0.0871	
Г	1234	1235			0.7290	0.6852		
	1245	1345			0.6532	0.4407		
Д	12345				0.7154			

Відносно позначень: наприклад, 1345 – означає лінійну регресію з чинниками x_1, x_3, x_4, x_5. З огляду отриманих результатів маємо: найкращою двофакторною моделлю є модель зі змінними x_1, x_2; трифакторною - x_1, x_2, x_3; чотирифакторною - x_1, x_2, x_3, x_4.

Далі порівнюємо моделі за $\hat{\sigma}_u^2$-критерієм: $\hat{\sigma}_u^2 = \frac{u^T u}{T - N}$. Найкращою моделлю в своїй групі є модель з найменшим $\hat{\sigma}_u^2$.

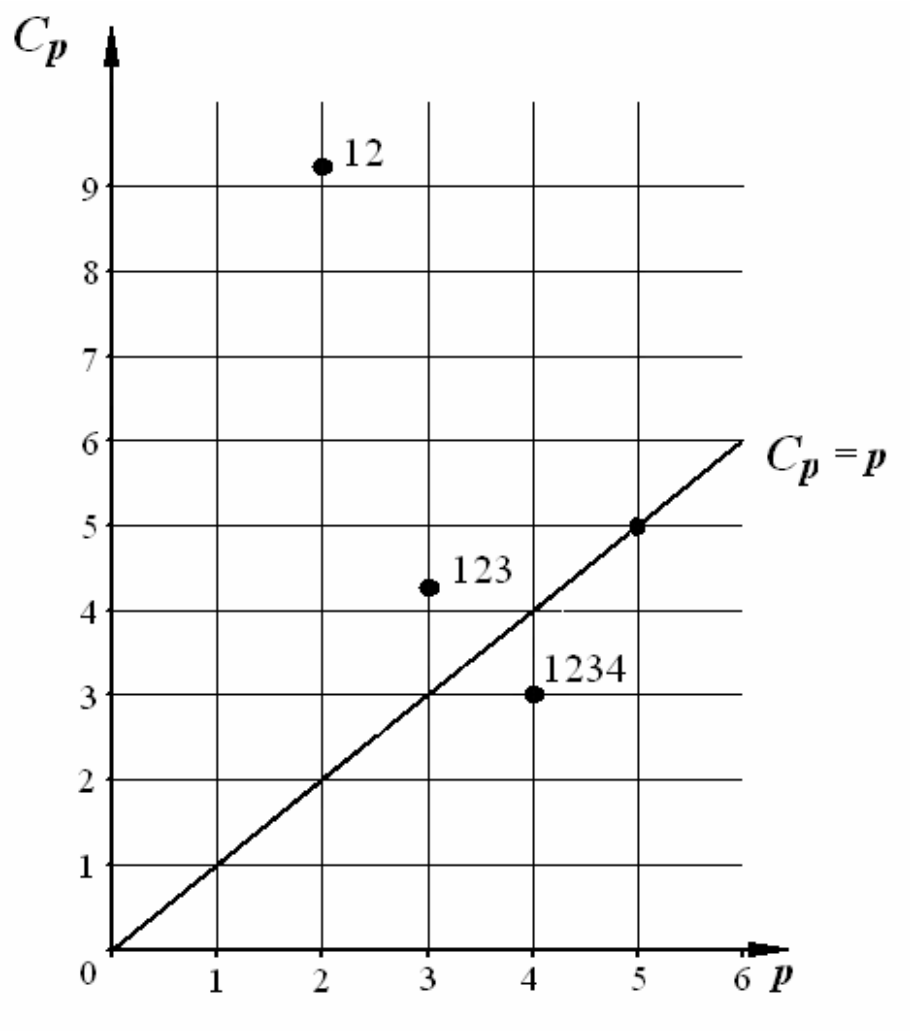
	Вид регресії				$\hat{\sigma}_u^2$			
Б	12	13	14	15	1.6817	2.6216	3.9425	4.6792
В	123	124	125		1.3512	1.4901	1.7546	
	134	135	145		2.4211	2.7223	4.1050	
Г	1234	1235			1.2188	1.4154		
	1245	1345			1.5593	2.5151		
Д	12345				1.2798			

Результати показують: найкращою двофакторною моделлю є модель x_1, x_2; трифакторною - x_1, x_2, x_3; чотирифакторною - x_1, x_2, x_3, x_4.

Залишилося зробити дослідження за допомогою C_p-статистики Маллоуза.

	Вид регресії				C_p-статистика			
Б	12	13	14	15	9.2238	26.1151	49.8538	63.0944
В	123	124	125		4.2276	6.6163	11.1623	
	134	135	145		22.6195	27.7973	51.5669	
Г	1234	1235			3	6.2255		
	1245	1345			8.5874	24.2707		
Д	12345				5			

Побудуємо графік залежності C_p-статистики від кількості чинників. Спочатку побудуємо пряму $C_p = p$. Вважається, що та модель краща, для якої C_p-статистика ближче знаходиться до прямої $C_p = p$.



Дивлячись на графік, можна зробити висновок, що найближче до прямої $C_p = p$ є точки, яким відповідають регресії $x_1\ x_2$; $x_1\ x_2\ x_3$; $x_1\ x_2\ x_3\ x_4$.

Аналізуючи результати трьох критеріїв, можемо зробити висновок, що «найкращими» в своїй групі є моделі зі змінними $x_1\ x_2$; $x_1\ x_2\ x_3$; $x_1\ x_2\ x_3\ x_4$ та сама довга. Далі економетрист, враховуючи мету побудови регресії, обирає з них одну «найкращу».

Будемо вважати, що для прогнозу ми обрали саму довгу модель

$$\hat{y} = 10.439\,x_1 - 14.755\,x_2 + 0.00003\,x_3 + 0.198\,x_4 - 0.00009\,x_5. \qquad (**)$$

Для з'ясування питання про специфікацію моделі скористаємося критерієм Рамсея.

Для моделі (**) розрахуємо $\hat{y}_t$ і u_{1t} ($t = \overline{1,25}$). Далі отримаємо, що $RSS_1 = \sum_{t=1}^{25} u_{1t}^2 = 25.5955$.

Будуємо модель

$$y^1 = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 \hat{y}^2 + \beta_7 \hat{y}^3 + \beta_8 \hat{y}^4 + u_2,$$

вона має вигляд

$$\hat{y}^1 = -65.6539 x_1 + 125.1472 x_2 - 0.0003 x_3 - 1.6972 x_4 - 0.0157 x_5 + \\ + 1.4473 \hat{y}^2 - 0.0948 \hat{y}^3 + 0.0023 \hat{y}^4.$$

Розраховуємо $\hat{y}_t^1$, u_{2t} і $RSS_2 = \sum_{t=1}^{25} u_{2t}^2$. Отримуємо: $RSS_2 = 24.0520$. Тепер можна розрахувати F_{cm}:

$$F_{cm} = \frac{(RSS_1 - RSS_2)(T-3)}{3RSS_2} = 0.4706.$$

$F_{kp} = F(\alpha; 3; T-3) = F(0.05; 3; 22) = 3.05$. Так як $F_{cm} < F_{kp}$, то обрана нами модель добре специфікована.

Скористаємося критерієм Амемії для знаходження найкраще специфікованої моделі з найкращих моделей у групі.

У групі Б найкращою за трьома критеріями є модель $\hat{y} = 14.1564 x_1 - 19.7675 x_2$. Значення вирішальної функції дорівнює $AF = 45.4065$.

У групі В: $\hat{y} = 11.7905 x_1 - 15.4081 x_2 + 0.00003 x_3$, $AF = 37.8331$.

Далі, у групі Г: $\hat{y} = 10.4377 x_1 - 14.7554 x_2 + 0.00003 x_3 + 0.1983 x_4$, $AF = 35.3462$.

І нарешті у групі Д:

$\hat{y} = 10.439 x_1 - 14.755 x_2 + 0.00003 x_3 + 0.198 x_4 - 0.00009 x_5$,
$AF = 38.3932$.

Так як у моделі з групи Г AF найменше, то модель $\hat{y} = 10.4377 x_1 - 14.7554 x_2 + 0.00003 x_3 + 0.1983 x_4$ краще всіх специфікована.

РОЗДІЛ 2. ОСОБЛИВІ ВИПАДКИ ПОБУДОВИ БАГАТОФАКТОРНОЇ РЕГРЕСІЙНОЇ МОДЕЛІ

Раніше ми будували і досліджували моделі при умові, що виконуються усі передумови 1МНК. Тепер розглядатимемо моделі, у яких умови 1МНК не виконуються. Ми розглянемо причини невиконання умов, наслідки, тестування наявності негативних явищ та їх усунення.

Розглянемо першу передумову: $M(u) = 0$ – математичне очікування залишків дорівнює нулеві.

Якщо ми маємо оцінену регресійну модель

$$\hat{y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_N x_N \text{ і } M(u) = a \neq 0,$$

то розглянемо нову регресійну модель

$$\hat{y} = (\hat{\beta}_1 - a) x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_N x_N,$$

у якій уже $M(u) = 0$.

ТЕМА 7. АВТОКОРЕЛЯЦІЯ ЗАЛИШКІВ

Будемо вважати, що друга передумова 1МНК не виконується. А це означає, що дисперсійно-коваріаційна матриця залишків матиме вигляд $\hat{\sum}_u = \hat{\sigma}_u^2 \Omega$, де матриця Ω уже не є одиничною.

Розглянемо випадок, коли недіагональні елементи матриці $\hat{\sum}_u$ не дорівнюють нулеві, тобто залишки u_t, $t = \overline{1,T}$ не є незалежними один від одного. У цьому випадку ми маємо справу з автокореляцією залишків.

Причини виникнення автокореляції залишків

Причинами появи автокореляції залишків можуть бути:

1) до регресійної моделі не включено чинники, що відіграють суттєву роль в економетричній моделі;
2) специфікація моделі виявилася невдалою;
3) при дослідженні економічного явища числові дані отримано з великими похибками;
4) інерційність (наявність лагу) та циклічність економічного процесу;
5) перетворення початкової специфікації моделі до лінійної форми.

Зауважимо, що автокореляція залишків викликає більш суттєву проблему у випадку, коли інтервал між спостереженнями досить малий. Отже, якщо інтервал між спостереженнями збільшити, то вплив неврахованих змінних буде зменшуватися.

Якщо кореляція між послідовними значеннями залишків додатна, то автокореляція називається додатною, а якщо від'ємна, то – від'ємною.

У економічних явищах від′ємна автокореляція зустрічається досить рідко. Вона може з'явитися при перетворенні нелінійної форми моделі до лінійної.

Наслідки автокореляції залишків

Якщо параметри моделі оцінювали методом 1МНК і у залишків існує автокореляція, то матимемо такі погані наслідки:

1) $\widehat{\boldsymbol{\beta}}$ не буде кращою оцінкою, тобто теорема Гауса – Маркова недійсна.
2) елементи коваріаційної матриці $\widehat{\sum}_{\widehat{\beta}}$ тенденційно «вниз» зміщені, а це негативно впливає на $\boldsymbol{t}$ - і $\boldsymbol{F}$ -тести, інтервали довіри і прогнозу (вони стають вузькими).

Авторегресійний процес першого порядку

Автокореляція залишків означає, що залишок $\boldsymbol{u_t}$ періоду $\boldsymbol{t}$ залежить від залишків більш ранніх періодів. Розглянемо простий вигляд залежності: $\boldsymbol{u_t}$ залежить від $\boldsymbol{u_{t-1}}$.

Означення. Залишки $\boldsymbol{u_t}$ задовольняють авторегресійному процесові першого порядку, якщо виконуються такі умови:

$$\boldsymbol{u_t = \rho u_{t-1} + \varepsilon_t}, \text{ де } \boldsymbol{|\rho| < 1,\ M(\varepsilon_t) = 0,\ t = \overline{1,T}}, \tag{1}$$

$$\boldsymbol{\sigma_{\varepsilon_{t'}\varepsilon_{t''}} = \begin{cases} 0, & t' \neq t'', \\ \sigma_\varepsilon^2, & t' = t''. \end{cases}} \tag{2}$$

Умови (1) і (2) гарантують, що вплив далеких від $\boldsymbol{u_t}$ залишків буде малим і випадкова змінна $\boldsymbol{\varepsilon_t}$ вільна від автокореляції і гетероскедастичності.

Авторегресійні процеси більш високого порядку можна висувати як альтернативну гіпотезу, якщо у початкових даних є циклічні коливання. Це може виникати при обробці даних по півріччям при наявності у них сезонних коливань. Тоді у наявності буде авторегресійний процес другого порядку.

Якщо ж ми маємо справу з квартальними даними, то можемо мати справу з авторегресійним процесом четвертого порядку.

Тестування автокореляції залишків

Найбільш відомим і поширеним тестом перевірки моделі на наявність автокореляції першого порядку є $\boldsymbol{d}$ -тест Дарбіна – Уотсона.

Спочатку зробимо декілька зауважень відносно застосування цього тесту:

1. $\boldsymbol{d}$ -тест Дарбіна – Уотсона виявляє лише автокореляцію залишків першого порядку. Для інших порядків використовуються інші тести.
2. Він не застосовується до моделей, у яких як незалежні змінні використовуються лагові значення залежної змінної. У цьому випадку використовують спеціальний тест $\boldsymbol{h}$ - Дарбіна.
3. Довжина вибірки не повинна бути менше 15.

Розглянемо етапи тестування автокореляції залишків тестом Дарбіна –Уотсона.

1. Формулюємо гіпотези:

$\boldsymbol{H_A : \rho \neq 0}$, присутня автокореляція залишків,

$H_0 : \rho = 0$, відсутня автокореляція залишків.

2. Задаємо рівень значущості α.
3. Розраховуємо значення d_{cm} за формулою

$$d_{cm} = \frac{\sum_{i=2}^{T} \left(u_i - u_{i-1} \right)^2}{\sum_{i=1}^{T} u_i^2}. \qquad (3)$$

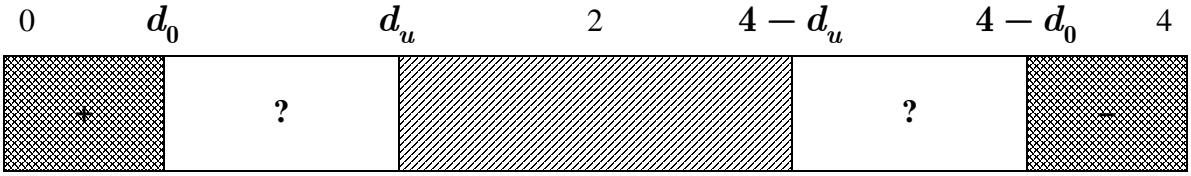
Можна показати, що при великих вибірках між d_{cm} і параметром ρ існує зв'язок:

$$d_{cm} \approx 2\left(1 - \rho\right).$$

Якщо автокореляція залишків відсутня, то $\rho = 0$ і d_{cm} наближається до двох. При наявності додатної автокореляції d_{cm} буде менша двох і наближатиметься до нуля, а при від′ємній – більша двох і наближатиметься до чотирьох. Отже, $d_{cm} \in \left[0;\ 4\right]$.

4. Критичне значення d при довільному рівні значущості залежить від кількості пояснюючих змінних у рівнянні регресії і від кількості спостережень у вибірці. Але воно також залежить від конкретних значень пояснюючих змінних. Тому неможливо скласти таблицю з точними критичними значеннями для всіх можливих вибірок. Однак, можна обчислити верхню і нижню межі для критичного значення d.

Далі представлено області прийняття рішень при d-тесті нульової гіпотези $H_0 : \rho = 0$.



– область відхилення гіпотези H_0,

– область невизначеності,

– область прийняття гіпотези H_0.

Якщо d_{cm} попадає в область невизначеності, то приймається гіпотеза про наявність автокореляції першого порядку.

Тест Дарбіна – Уотсона має певні недоліки: наявність зони невизначеності, визначає тільки автокореляцію першого порядку. Тому розглянемо і інші тести.

Тест серій (Бреуша – Годфрі)

Ідея тесту дуже проста і полягає в тому, що якщо існує автокореляція залишків першого порядку, то у рівнянні

$$u_t = \rho u_{t-1} + \varepsilon_t,\ t = \overline{1,T}$$

коефіцієнт ρ буде значимо відрізнятися від нуля. Цей коефіцієнт оцінюємо 1МНК.

Перевага тесту Бреуша – Годфрі в порівнянні з тестом Дарбіна – Уотсона полягає в тому, що він перевіряється за допомогою статистичного критерію, а тест Дарбіна – Уотсона має зону невизначеності для значень статистики d. Другою перевагою тесту є можливість включати до регресорів залишки не тільки з лагом 1, а і з лагами 2, 3 і т. д. Це дозволяє виявити автокореляцію більш високих порядків.

Наприклад, розглянемо оцінену авторегресійну залежність залишків вигляду

$$\underset{(0.10)}{u_t = 0.61\,u_{t-1}} \underset{(0.11)}{- 0.15\,u_{t-2}} + \underset{(0.10)}{0.01\,u_{t-3}},$$

Внизу у дужках виписані стандартні відхилення коефіцієнтів. Зрозуміло, що значущим буде тільки перший коефіцієнт ($t_{cm} = 6.1$). Це означає, що існує авторегресійний процес першого порядку.

Тест на автокореляцію залишків для моделей з лаговою залежною змінною

При розгляді тесту Дарбіна – Уотсона ми записали, що його неможливо застосовувати у присутності в рівнянні лагових залежних змінних. Тому Дарбін запропонував новий h-тест Дарбіна.

Нехай регресійне рівняння має вигляд

$$y = \alpha + \beta_1 x + \beta_2 y_{t-1} + u_t.$$

Далі обчислюється h_{cm}:

$$h_{cm} = \hat{\rho}\sqrt{\frac{n}{1 - T\hat{\sigma}^2_{\beta_2}}},$$

де $\hat{\rho}$ – оцінка ρ в означенні автокореляції першого порядку. Можна брати $\hat{\rho} \approx \left(1 - 0.5\,d_{cm}\right)$, де d_{cm} – оцінка статистики Дарбіна – Уотсона, T – довжина вибірки.

Розглядається гіпотеза: $H_0 : \rho = 0$. Вона відхиляється, якщо $h_{cm} > 1.96$ $\left(\alpha = 0.05\right)$ або $h_{cm} > 2.58$ $\left(\alpha = 0.01\right)$.

Зазначимо, що тест працює при великій вибірці.

Методи усунення автокореляції

Друга передумова 1МНК має вигляд: $\sum_u = \sigma_u^2 I$. Якщо вона не виконується, то тоді $\widehat{\sum}_u = \sigma_u^2 \Omega$, і модель називається узагальненою регресійною. Якщо матриця Ω нам відома, то параметри моделі можна оцінити за допомогою Узагальненого методу найменших квадратів (УМНК). Узагальнена оцінка методом УМНК називається також оцінкою Ейткена і розраховується таким чином:

$$\widetilde{\beta} = \left(X^T \Omega^{-1} X\right)^{-1} X^T \Omega^{-1} Y,$$
$$\widetilde{\sum}_{\widetilde{\beta}} = \widetilde{\sigma}_u^2 \left(X^T \Omega^{-1} X\right)^{-1},$$
$$\widetilde{\sigma}_u^2 = \frac{\widetilde{u}^T \Omega^{-1} \widetilde{u}}{T - N}, \qquad \text{(A)}$$

$$\widetilde{u} = y - X\widetilde{\beta}.$$

Якщо існує авторегресійний процес першого порядку, то матриця Ω має вигляд:

$$\Omega = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \dots & \rho^{T-1} \\ \rho & 1 & \rho & \rho^2 & \dots & \rho^{T-2} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \rho^{T-4} & \dots & 1 \end{pmatrix}.$$

Знак $(\sim)$ означає оцінку методом УМНК.

Оцінка узагальненої регресійної моделі, що опирається на оцінене значення $\hat{\rho}$, буде оцінкою, подібною оцінці Ейткена. Алгоритм має такий вигляд:

1. Розрахунок вектора помилок при оцінці методом 1МНК.
2. Використовуємо тест Дарбіна – Уотсона на наявність авторегресійного процесу першого порядку. Якщо автокореляція існує, то переходимо до пункту 3, а якщо ні, то процес закінчено.
3. Розраховуємо оцінку параметра ρ за допомогою формули:

$$\hat{\rho} = \frac{\sum_{t=2}^{T} u_t u_{t-1}}{\sum_{t=1}^{T} u_t^2}.$$

4. Матрицю початкових даних $D = \left[Y \mid X\right]$ перетворюємо за допомогою матриці перетворень $T^A : T^A D = \left[T^A Y \mid T^A X\right] = D^* = \left[Y^* \mid X^*\right]$. Якщо залишки

u_t задовольняють авторегресійному процесу першого порядку, то матриця T^A має вигляд:

$$T^A = \begin{pmatrix} \sqrt{1-\hat{\rho}^2} & 0 & 0 & \dots & 0 & 0 \\ -\hat{\rho} & 1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & -\hat{\rho} & 1 \end{pmatrix}.$$

5. До перетворених початкових даних D^* застосовуємо метод 1МНК:

$$\widetilde{\beta} = \left(X^{*T}X^*\right)^{-1} X^{*T}Y^*,$$

$$\widetilde{\sum}_{\widetilde{\beta}} = \widetilde{\sigma}^2_{\widetilde{u}^*} \left(X^{*T}X^*\right)^{-1},$$

$$\widetilde{\sigma}^2_{\widetilde{u}^*} = \frac{\widetilde{u}^{*T}\widetilde{u}^*}{T-N}, \quad \widetilde{u}^* = y^* - \widetilde{y}^*.$$

Перша формула дає нам Ейткен-оцінку для β, а не 1МНК-оцінку. Оцінки $\widetilde{\beta}, \widetilde{\sum}_{\widetilde{\beta}}, \widetilde{\sigma}^2_{\widetilde{u}^*}, \widetilde{u}^*$ будуть такими ж, як і за формулами (А). Якщо параметр ρ відомий, то оцінка Ейткена в узагальненій регресійній моделі має оптимальні властивості, як і 1МНК-оцінка в класичній моделі, тобто справедлива теорема Гауса – Маркова.

Якщо авторегресійний параметр ρ невідомий, то про статистичні властивості Ейткен-подібних оцінок для β (тут ми їх назвали також Ейткен-оцінками) при малій довжині ряду даних T важко щось сказати певне. Тому ми в таких випадках кажемо про асимптотичні властивості, що притаманні рядам даних великої довжини.

Оцінка моделі, якщо параметр ρ невідомий

Ситуації, коли параметр авторегресії ρ відомий, зустрічаються дуже рідко. Тому виникає необхідність у процедурах оцінення при невідомому ρ. Як правило, вони мають ітеративний характер.

Метод Кохрейна – Оркатта

Розглядається регресійна модель

$$y = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_N x_N + u.$$

Спочатку оцінимо коефіцієнти моделі 1МНК і розрахуємо залишки u_t, $t = \overline{1,T}$. Далі

1. За початкове наближення ρ візьмемо його 1МНК-оцінку $\hat{\rho}$ в регресії $u_t = \rho u_{t-1} + \varepsilon_t$, тобто

$$\hat{\rho} = \frac{\sum_{t=2}^{T} u_t\, u_{t-1}}{\sum_{t=1}^{T} u_t^2}.$$

2. Робимо перетворення початкових даних за допомогою матриці T^A

$$T^A = \begin{pmatrix} \sqrt{1-\hat{\rho}^2} & 0 & 0 & \dots & 0 & 0 \\ -\hat{\rho} & 1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & -\hat{\rho} & 1 \end{pmatrix}$$

і знаходимо 1МНК-оцінки $\widehat{\beta}$ параметра β.

3. Будуємо новий вектор залишків $u = y - X\widehat{\beta}$.
4. Переходимо до пункту 1.

Метод Дарбіна

Розглянемо регресію

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \dots + \beta_N x_{Nt} + u_t. \qquad (1)$$

Випишемо рівняння при $t := t-1$:

$$y_{t-1} = \beta_1 x_{1,t-1} + \beta_2 x_{2,t-1} + \dots + \beta_N x_{N,t-1} + u_{t-1}. \qquad (2)$$

Домножимо (2) на (ρ) і віднімемо від (1):

$$y_t - \rho\, y_{t-1} = \beta_1\left(x_{1t} - \rho\, x_{1,t-1}\right) + \beta_2\left(x_{2t} - \rho\, x_{2,t-1}\right) + \dots + \\ + \beta_N\left(x_{Nt} - \rho\, x_{N,t-1}\right) + u_t - \rho\, u_{t-1}.$$

А так як присутній авторегресійний процес першого порядку, то:

$$u_t = \rho u_{t-1} + \varepsilon_t, \qquad |\rho| < 1,\ M(\varepsilon_t) = 0,\ t = \overline{1,T},$$

$$\sigma_{\varepsilon_{t'}\varepsilon_{t''}} = \begin{cases} 0, & t' \neq t'', \\ \sigma_\varepsilon^2, & t' = t''. \end{cases}$$

Звідси маємо, що $u_t - \rho\, u_{t-1} = \varepsilon_t$, де залишки ε_t не мають авторегресії. Отже, отримали таку модель:

$$y_t = \rho\, y_{t-1} + \beta_1 x_{1t}\left(1-\rho\right) + \beta_2\left(x_{2t} - \rho\, x_{2,t-1}\right) + \dots + \\ + \beta_N\left(x_{Nt} - \rho\, x_{N,t-1}\right) + \varepsilon_t,$$

яка вільна від автокореляції залишків.

Далі застосовуємо 1МНК для оцінки коефіцієнтів цієї моделі. Отримаємо оцінку $\hat{\rho}$ (коефіцієнт при y_{t-1}). На наступному кроці оцінка $\hat{\rho}$ використовується для розрахунків перетворених змінних $\left(y_t - \hat{\rho}\, y_{t-1}\right)$ і $\left(x_t - \hat{\rho}\, x_{t-1}\right)$. Знову до нових змінних застосовуємо 1МНК. Тоді коефіцієнти при $\left(x_{it} - \hat{\rho}\, x_{i,t-1}\right)$ будуть оцінками $\hat{\beta}_i$, а коефіцієнт при x_{1t}, поділений на $\left(1 - \hat{\rho}\right)$, дає $\hat{\beta}_1$.

Зазначимо, що перетворення змінних можна робити за допомогою перетворення T^A:

$$T^A = \begin{pmatrix} -\rho & 1 & 0 & \dots & 0 & 0 \\ 0 & -\rho & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & -\rho & 1 \end{pmatrix},$$ порядок матриці $T^A : \left(T-1\right) \times T$.

Тоді перетворення робимо за правилом: $\left[T^A Y \mid T^A X\right]$. Тут ми губимо перше спостереження: $y_1, x_{11}, x_{21}, \dots, x_{N1}$.

Приклад 10. Для моделі А необхідно:

1) дослідити залишки на автокореляцію першого порядку за допомогою тесту Дарбіна – Уотсона;
2) при наявності автокореляції залишків першого порядку оцінити коефіцієнти моделі за допомогою методу Кохрейна – Оркатта.

Нагадаємо, що модель А має вигляд:

$$\hat{y} = 10.439\, x_1 - 14.755\, x_2 + 0.00003\, x_3 + 0.198\, x_4 - 0.00009\, x_5.$$

Формулюємо гіпотези: $H_A : \rho \neq 0$, $H_0 : \rho = 0$. Візьмемо $\alpha = 0.05$. За формулою (3) розраховуємо d_{cm}:

$$d_{cm} = \frac{\sum_{i=2}^{T} \left(u_i - u_{i-1}\right)^2}{\sum_{i=1}^{T} u_i^2} = 1.6368.$$

За таблицями Дарбіна – Уотсона при $\alpha = 0.05$, $T = 25$, $N = 5$ знаходимо: $d_0 = 1.038$, $d_u = 1.767$. Сформуємо таблицю

0	$d_0 = 1.038$	$d_u = 1.767$	2	$4 - d_u = 2.233$	$4 - d_0 = 2.962$	4
+	?			?	−	

Ми бачимо, що $d_0 < d_{cm} < d_u$, тобто, d_{cm} попадає в область невизначеності. Це означає, що нічого не можна сказати про наявність або відсутність автокореляції залишків першого порядку. Тому вважаємо, що автокореляція залишків першого порядку присутня з ймовірністю 95 %.

Усунемо автокореляцію залишків методом Кохрейна – Оркатта. Розрахуємо параметр $\hat{\rho}$:

$$\hat{\rho} = \frac{\sum_{i=2}^{T} u_i u_{i-1}}{\sum_{i=1}^{T} u_i^2} = 0.1756.$$

Тепер побудуємо матрицю перетворень T^A (вона має вимір 25×25):

$$T^A = \begin{pmatrix} \sqrt{1-\hat{\rho}^2} & 0 & 0 & \dots & 0 & 0 \\ -\rho & 1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \ddots & \dots & \dots \\ 0 & 0 & 0 & \dots & -\rho & 1 \end{pmatrix}.$$

Далі перетворимо початкові дані $D = \left[Y \mid X \right]$ за допомогою матриці перетворень T^A: $T^A D = \left[T^A Y \mid T^A X \right] = \left[Y^1 \mid X^1 \right]$ і до нових даних застосуємо метод 1МНК. Отримаємо модель

$$\text{A: } \hat{y} = 10.712\, x_1 - 14.428\, x_2 + 0.00003\, x_3 + 0.229\, x_4 - 0.026\, x_5. \qquad (4)$$

Розрахуємо залишки і оцінимо за формулою (3) d_{cm}:

$$d_{cm} = 1.8775.$$

На цей раз d_{cm} попадає в область прийняття гіпотези H_0, а це означає, що у отриманої моделі (4) автокореляція залишків першого порядку вже відсутня.

ТЕМА 8. ГЕТЕРОСКЕДАСТИЧНІСТЬ ЗАЛИШКІВ

Друга передумова 1МНК передбачає однакову дисперсію у залишків u_t, $t = \overline{1,T}$. Таке твердження може здаватися дивним, адже у кожному спостереженні випадковий член має тільки одне значення, і тому виникає питання, а що тоді означає його дисперсія.

Мається на увазі його можлива поведінка до того, як буде зроблена вибірка. Тобто, ймовірність того, що величина u_t прийме якесь значення, буде однаковою для всіх спостережень. Ця умова носить назву «гомоскедастичність», що означає «однаковий розкид».

Разом з тим для деяких вибірок більш доцільно вважати, що теоретичний розподіл випадкового члена є різним для різних спостережень у вибірці. Тобто апріорна ймовірність отримання великих відхилень буде досить великою. Це приклад гетероскедастичності, що означає «неоднаковий розкид».

Означення. Явище, коли дисперсії залишків неоднакові, називається гетероскедастичністю залишків, тобто $\sigma^2_{u_i} \neq \sigma^2_u$ хоча б для деяких i, $i = \overline{1,T}$, тобто

$$\Sigma_u = \sigma_u^2 \Omega = \sigma_u^2 \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_T^2 \end{pmatrix}.$$

Причини виникнення гетероскедастичності залишків

Гетероскедастичність залишків може виникнути, якщо:

1) об’єкти дослідження неоднорідні, тобто ми досліджуємо разом великі і малі об’єкти;
2) спостереження є членами часового ряду і якщо регресори і регресанд одночасно збільшуються або зменшуються;
3) невключені змінні і похибки виміру, впливаючи на залишок, роблять його досить малим при малих значеннях регресорів і регресанда і досить великим при їх значних значеннях;
4) невдалою є специфікація моделі.

Наслідки гетероскедастичності залишків

Наслідки гетероскедастичності залишків можуть бути такі:

1) зменшується ефективність оцінок коефіцієнтів моделі, тобто можна знайти інші оцінки, що матимуть менші дисперсії і будуть незміщеними;
2) стандартні помилки будуть заниженими, а тому буде неправильне уявлення про точність оцінок рівняння регресії, інтервали довіри, значущість оцінок.

Методи тестування гетероскедастичності залишків

Розглянемо декілька загальновживаних тестів на виявлення гетероскедастичності залишків моделі

$$y = \sum_{i=1}^{N} \beta_i x_i + u.$$

Усі відомі тести можна розбити на дві групи. До першої групи відносяться тести Спірмена і Гольдфельда – Квандта, які дозволяють визначити наявність або відсутність гетероскедастичності, але вони не дають можливості дослідити кількісний характер залежності дисперсій помилок регресії від значень регресорів і тому не допомагають усунути гетероскедастичність залишків.

Друга група – це достатні умови існування гетероскедастичності. Це означає, що якщо тест не виконується, то не можна вважати, що гетероскедастичність відсутня. Ця група тестів допомагає усунути гетероскедастичність залишків.

Тест Гольдфельда – Квандта

Цей тест можна застосовувати як до малих, так і до великих вибірок, якщо справедливе припущення про пряму залежність дисперсії похибки від величини деякої незалежної змінної. Також необхідно, щоб помилки регресії були нормально розподіленими випадковими величинами.

Якщо ми не знаємо, від якої незалежної змінної залежать дисперсії похибки, то найбільш впливову змінну і обираємо. А ще краще застосувати тест до кожної із змінних (крім x_1).

Алгоритм тесту Гольдфельда – Квандта

Нехай дисперсія похибок залежить від змінної x_p. Тоді:

1) упорядковують початкові дані по мірі росту незалежної змінної x_p, відносно якої є підозра на гетероскедастичність;
2) видаляють m середніх спостережень ($m = \frac{4}{15}T$);
3) на основі значень кожної із двох груп спостережень будують відповідно дві регресійні моделі методом 1МНК і розраховують залишки для кожної із них:

$$D = [Y \mid X] = \begin{bmatrix} \underline{Y1} & | & \underline{X1} \\ & | & \\ \overline{Y2} & | & \overline{X2} \end{bmatrix} \begin{matrix} T_1 \times (N+1) \\ m \\ T_2 \times (N+1) \end{matrix},$$

$$T_1 + T_2 + m = T.$$

4) розраховують F_{cm}:

$$F_{cm} = \frac{\sum_{t=1}^{T_2} u_{2t}^2 \Big/ (T_2 - N)}{\sum_{t=1}^{T_1} u_{1t}^2 \Big/ (T_1 - N)},$$

де u_{kt}, $k = 1, 2$ – залишки k-ї групи спостережень;

5) якщо $F_{cm} > F_{kp} = F\left(\alpha, T_2 - N, T_1 - N\right)$, то гетероскедастичність залишків існує з ймовірністю $\left(1-\alpha\right)100\%$.

Зауважимо, що цей тест працює і без виключення m середніх спостережень, але при цьому його потужність зменшується.

Тест рангової кореляції Спірмена

Тест Спірмена застосовується як до малих, так і до великих вибірок при умові, що дисперсія похибок лінійно залежить від значень змінної x_p. Якщо ж залежність нелінійна, то доцільно користуватися тестом Кендалла.

Алгоритм тесту Спірмена

1) розраховуємо залишки моделі;
2) ранжуємо $\left|u_{pt}\right|$ та x_{pt}, $t = \overline{1,T}$ у зростаючому чи спадному порядку;
3) розраховуємо коефіцієнт рангової кореляції Спірмена за формулою

$$r^C = 1 - \frac{6\sum_{i=1}^{T} d_i^2}{T\left(T^2-1\right)},$$

де d_i – різниця між рангами, що приписуються $\left|u_{pt}\right|$ і x_{pt};

4) перевіряємо значущість коефіцієнта Спірмена за критерієм Ст'юдента. Для цього розраховуємо t-статистику:

$$t_{cm} = \frac{\left|r^C\right|\sqrt{T-2}}{\sqrt{1-\left(r^C\right)^2}};$$

5) якщо $t_{cm} > t_{kp} = t\left(\frac{\alpha}{2}, T-2\right)$, то коефіцієнт Спірмена значимо відрізняється від нуля з ймовірністю $\left(1-\alpha\right)100\%$.

Відомо, що $\left|r^C\right| \le 1$. Якщо $\left|r^C\right| \ge 0.6$, то вважаємо, що гетероскедастичність залишків існує.

Розглянемо другу групу тестів, що дозволяють усунути гетероскедастичність, якщо вона присутня.

Тест Парка

Алгоритм його такий:

1) за допомогою 1МНК оцінюємо коефіцієнти регресії і розраховуємо залишки u_t;
2) будуємо допоміжну модель

$$\ln\left|u\right| = b_1 z_1 + b_2 \ln\left|z_2\right| + \varepsilon, \qquad (1)$$

де $z_{1t} \equiv 1,\ t = \overline{1,T}$, а z_2 – незалежна змінна, відносно якої ми вважаємо, що існує гетероскедастичність залишків;

3) за допомогою 1МНК оцінюємо коефіцієнти b_1 і b_2;

4) перевіряємо на значущість коефіцієнти b_1 і b_2 за допомогою t-критерія Ст'юдента (див. тему 4). Якщо b_1 і b_2 значимо відрізняються від нуля, то існує змішана гетероскедастичність, а якщо b_1 незначимо відрізняється від нуля, а b_2 – значимо, то кажуть, що існує чиста гетероскедастичність залишків. Коли ж b_1 і b_2 незначимо відрізняються від нуля, то залежності (1) не існує між залишками і регресором x_p, і тому необхідно шукати іншу залежність, тобто користуватися іншим тестом.

Тест Уайта

Цей тест досить часто використовується. Будемо вважати, що дисперсія похибок регресії є одна і та ж функція від спостережених значень регресорів, тобто:

$$\sigma_t^2 = f\left(x_t\right),\ t = \overline{1,T}. \tag{2}$$

Досить часто функцію $f\left(x\right)$ обирають квадратичною, що відповідає тому, що середня квадратична похибка регресії залежить від спостережених значень регресорів приблизно лінійно.

Уайт пропонує оцінювати (2) за допомогою квадратів залишків:

$$u_t^2 = f\left(x_t\right) + \varepsilon_t,\ t = \overline{1,T}, \tag{3}$$

де ε_t – випадковий член. Якщо регресія (3) значимо відрізняється від нуля, то гетероскедастичність залишків існує.

Тест Глейзера

Цей тест схожий на тест Уайта. Тільки у якості функції $f\left(x\right)$ обирається функція

$$f\left(x\right) = \alpha + \gamma x_p^s$$

і розглядається регресія

$$\left|u_t\right| = \alpha + \gamma x_{pt}^s + \varepsilon_t, \tag{4}$$

де $s = 1,\ 2,\ \ldots$ Регресія (4) оцінюється при різних значеннях s. Серед регресій, що значимо відрізняються від нуля, обираємо ту, що дає максимальну t_{cm} коефіцієнта γ.

Методи усунення гетероскедастичності залишків

Вважаємо, що автокореляція залишків у моделі відсутня.

Якщо встановлено за будь-яким тестом наявність гетероскедастичності, то початкову модель перетворюємо так, щоб помилки мали постійну дисперсію. Потім невідомі параметри трансформованої моделі оцінюються за методом 1МНК. Вид перетворення первинної моделі

залежить від форми гетероскедастичності залишків, тобто від форми залежності між дисперсією залишків та значеннями незалежних змінних.

Матриця перетворення T^H при наявності гетероскедастичності залишків має вигляд:

$$T^H = \begin{pmatrix} \frac{1}{\sigma_1} & 0 & 0 & \dots & 0 & 0 \\ 0 & \frac{1}{\sigma_2} & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \frac{1}{\sigma_{T-1}} & 0 \\ 0 & 0 & 0 & \dots & 0 & \frac{1}{\sigma_T} \end{pmatrix},$$

де σ_i, $i = \overline{1,T}$ – стандартні відхилення залишків. Величини σ_i можуть бути невідомі. Визначати матрицю T^H можна різними способами:

1) без використання статистичної оцінки;
2) визначення T^H за принципом «гетероскедастичність між гомоскедастичними групами»;
3) σ_i або σ_i^2 можна отримати як оцінки $|u_i|$ і u_i^2 у регресіях (1), (4) і (3);
4) можна скористатися узагальненим методом найменших квадратів. Метод, що використовує матрицю T^H, називають зваженим методом найменших квадратів.

1. Визначення T^H без статистичної оцінки.

У цьому випадку достатньо зовнішньої інформації, щоб без статистичної оцінки визначити діагональні елементи матриці перетворень T^H. Якщо відомо, що $\sigma_i = a x_{ki}$ або $\sigma_i^2 = a x_{ki}$, то елементи матриці T^H такі: $\frac{1}{a x_{ki}}$ або $\frac{1}{\sqrt{a x_{ki}}}$, $i = \overline{1,T}$. Константі a можна надати значення «1». Це не впливає на результат.

2. Оцінка T^H за принципом «гетероскедастичність між гомоскедастичними групами».

Досить часто зовнішньої інформації недостатньо, щоб визначити діагональні елементи матриці T^H без статистичної оцінки. Тому ці елементі повинні бути статистично оцінені. Насправді оцінка діагональних елементів матриці T^H зводиться до розгляду гетероскедастичності між гомоскедастичними групами.

Наприклад, якщо $\sigma_{u_t}^2 = \sigma_1^2$ для $x_{kt} < 20$ – перша група, $\sigma_{u_t}^2 = \sigma_2^2$ для $x_{kt} \geq 20$ – друга група, $1 \leq k \leq N$, це означає, що в групах існує гомоскедастичність, а між групами – гетероскедастичність. У цьому випадку необхідно оцінити σ_1^2 і σ_2^2. Оцінки

можна робити, як у тесті Гольдфельда – Квандта. Тоді діагональними елементами матриці T^H будуть:

$$\underbrace{\frac{1}{\hat{\sigma}_1}, \frac{1}{\hat{\sigma}_1}, \ldots, \frac{1}{\hat{\sigma}_1}}_{\text{для групи 1}}, \underbrace{\frac{1}{\hat{\sigma}_2}, \frac{1}{\hat{\sigma}_2}, \ldots, \frac{1}{\hat{\sigma}_2}}_{\text{для групи 2}}.$$

3. На практиці значення σ_i майже ніколи не бувають відомі. У цьому випадку можна скористатися таким алгоритмом:

1) оцінюємо модель методом 1МНК і розраховуємо залишки;
2) будуємо одну з моделей (1), (3), (4) або досить часто розглядають модель (2), де $f(x)$ є квадратичною функцією регресорів;
3) розраховуємо $\left|\hat{u}_i\right|$ або $\hat{u}_i^2$, $i = \overline{1,T}$, ці оцінки і будуть оцінками σ_i або σ_i^2, $i = \overline{1,T}$.

4. Якщо існує гетероскедастичність залишків, то матриця Σ_u має вигляд:

$$\Sigma_u = \sigma_u^2 \Omega = \sigma_u^2 \begin{pmatrix} \sigma_1^2 & 0 & \ldots & 0 & 0 \\ 0 & \sigma_2^2 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 & \sigma_T^2 \end{pmatrix}.$$

Якщо σ_i^2 відомі, то можна скористатися узагальненим методом найменших квадратів, наприклад, $\tilde{b} = \left(X^T \Omega^{-1} X\right)^{-1} X^T \Omega^{-1} Y$.

Приклад 11. Перевірити на гетероскедастичність залишків модель А, яку звільнено від автокореляції залишків:

$$\hat{y} = 10.712\, x_1 - 14.428\, x_2 + 0.00003\, x_3 + 0.229\, x_4 - 0.026\, x_5.$$

Так як ми не знаємо, відносно якої змінної може існувати гетероскедастичність залишків, то ми застосовуємо тест Гольдфельда – Квандта, використовуючи послідовно усі змінні: x_2, x_3, x_4, x_5. Якщо відносно декількох змінних існує підозра на гетероскедастичність, то обираємо ту, для якої F_{cm} найбільша.

У нашому прикладі відносно кожної із змінних x_2, x_3, x_4, x_5 упорядковуємо початкові дані по мірі зросту кожної із змінних. Видаляємо $m = \frac{4}{15}T = \frac{4}{15} \cdot 25 \approx 7$

середніх спостережень, і для кожної із отриманих двох груп будуємо регресійні моделі, розраховуємо залишки і F_{cm}.

У нашому прикладі найбільшу статистику має змінна x_4. Опишемо алгоритм побудови статистики для змінної x_4. Так як $m = 7$, то після видалення семи спостережень залишиться тільки 25 – 7 = 18 спостережень. Розбиваємо їх на дві групи по 9 спостережень у порядку збільшення значень змінної x_4.

	y	x_1	x_2	x_3	x_4	x_5	
1	5.88	1	0.49	25587	3.46	23.99	Перша група спостережень
2	6.22	1	0.37	19459	3.56	25.68	
3	6.30	1	0.36	16821	3.60	21.76	
4	8.72	1	0.26	18946	4.25	13.69	
5	6.50	1	0.35	50907	4.28	25.74	
6	9.37	1	0.23	45893	4.36	15.98	
7	8.17	1	0.31	52509	4.50	21.92	
8	7.02	1	0.32	20068	4.82	13.21	
9	9.12	1	0.26	14903	4.88	19.52	
17	7.37	1	0.30	26705	7.19	16.38	Друга група спостережень
18	9.38	1	0.24	50391	7.80	18.39	
19	10.81	1	0.17	41089	7.90	22.37	
20	4.32	1	0.42	5736	8.52	22.97	
21	6.61	1	0.38	6920	8.85	21.21	
22	5.52	1	0.34	11049	9.27	20.09	
23	12.11	1	0.19	43149	9.76	26.46	
24	9.87	1	0.43	22661	9.90	17.55	
25	13.17	1	0.17	99400	10.31	18.27	

За даними двох груп будуємо дві моделі і розраховуємо залишки для цих моделей:

$$\hat{y}^1 = 11.763\,x_1 - 13.4637\,x_2 + 0.000007\,x_3 + 0.1378\,x_4 - 0.0325\,x_5,$$

$$\hat{y}^2 = 5.3926\,x_1 - 15.5122\,x_2 + 0.00003\,x_3 + 0.9299\,x_4 - 0.0651\,x_5.$$

$$\hat{u}^1 = \begin{pmatrix} 0.8383 \\ -0.3533 \\ -0.5225 \\ 0.1844 \\ -0.6597 \\ 0.3013 \\ 0.3060 \\ -0.8098 \\ 0.7153 \end{pmatrix}, \qquad \hat{u}^2 = \begin{pmatrix} 0.1757 \\ 0.0775 \\ 0.8789 \\ -1.1645 \\ 0.0465 \\ -2.2566 \\ 0.9613 \\ 2.3749 \\ -1.0937 \end{pmatrix}.$$

Розраховуємо F_{cm} : $F_{cm} = \dfrac{\sum_{i=1}^{9}\left(u_i^2\right)^2 / \left(T_2 - N\right)}{\sum_{i=1}^{9}\left(u_i^1\right)^2 / \left(T_1 - N\right)} = 5.1407$. Так як $F_{cm} > F_{kp} =$

$= F\left(0.05;\ 9-4;\ 9-4\right) = F\left(0.05;\ 5;\ 5\right) = 5.05$, то гетероскедастичність залишків існує з ймовірністю 95 %.

Так як доведено, що гетероскедастичність залишків існує, то тепер необхідно розрахувати $\sigma_i\left(i = \overline{1,25}\right)$ для матриці перетворень T^H. Скористаємося третім способом розрахунку σ_i.

Вважаємо, що існує така залежність:

$$u_t^2 = a_1 x_{1t} + a_2 x_{2t} + a_3 x_{3t} + a_4 x_{4t} + a_5 x_{5t} + \varepsilon_t .$$

Коефіцієнти цієї регресії оцінимо методом 1МНК:

$$\hat{u}_t^2 = -4.8967\, x_{1t} + 8.2364\, x_{2t} - 0.00001\, x_{3t} + 0.4591\, x_{4t} + 0.0495\, x_{5t} .\ (*)$$

Якщо доведемо, що ця модель значимо відрізняється від нуля, то це буде підтвердженням того, що гетероскедастичність залишків існує, і між u_t^2 і x_i $(i = \overline{1,5})$ існує залежність (*).

Знаходимо прогнозні значення $\hat{u}_t^2$ і залишки $\hat{\varepsilon}_t$:

$$\hat{u}_t^2 = \Big(0.2799, 0.9634, 1.8028, 0.4347, 0.5506, 3.6867, -0.4423, 0.2464, 1.5814,$$
$$0.3253, 0.7332, 1.8447, 0.3116, 3.1527, 2.9731, 0.7672, 0.1393, 0.1911,$$
$$0.7652, -0.4787, 1.1940, 0.5052, 2.9074, -1.2500, 1.1669\Big)^T ,$$

$$\varepsilon_t = \Big(-0.0124, -0.6628, 0.2966, -0.4234, 0.4870, 3.6528, 0.7613,\ 1.0125,\ -0.3158,$$
$$-0.2843,\ -0.7184,\ -1.7855,\ 0.5283,\ -3.1353,\ 0.2860,\ -0.2981,\ -0.0806,$$
$$-0.1910,\ -0.7588,\ 0.8455,\ -0.9897,\ -0.5006,\ 1.9871,\ 1.4510,\ -1.1492\Big)^T .$$

За ними розраховуємо $F_{cm} = \dfrac{\hat{u}^{2T}\hat{u}^2\left(T-N\right)}{\hat{\varepsilon}^T\hat{\varepsilon}\left(N-1\right)} = 22.5571$. Розраховуємо

$F_{kp} = F\left(\alpha; N-1; T-N\right) = F\left(0.05; 5-1; 25-5\right) = 2.87$. Так як $F_{cm} >$

$> F_{kp}$, то з ймовірністю 95 % модель (*) значимо відрізняється від нуля. Ми ще раз довели, що гетероскедастичність залишків існує.

Для усунення гетероскедастичності залишків оцінимо σ_i: $\sigma_i = \sqrt{\hat{u}_i^2}$.

$$T^H = \begin{pmatrix} \frac{1}{\sqrt{0.2799}} & 0 & 0 & \dots & 0 & 0 \\ 0 & \frac{1}{\sqrt{0.9634}} & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \frac{1}{\sqrt{1.2500}} & 0 \\ 0 & 0 & 0 & \dots & 0 & \frac{1}{\sqrt{1.1669}} \end{pmatrix}.$$

Початкові дані $\left[Y \mid X \right]$ перетворюємо за допомогою матриці перетворень T^H: $\left[T^H Y \mid T^H X \right] = \left[Y^1 \mid X^1 \right]$. Далі використовуємо $\left[Y^1 \mid X^1 \right]$ для побудови регресійної моделі методом 1МНК, у якої гетероскедастичність залишків уже відсутня:

$$\hat{y} = 11.7747\, x_1 - 18.7448\, x_2 + 0.00001\, x_3 + 0.1784\, x_4 + 0.0216\, x_5 .$$

ТЕМА 9. МУЛЬТИКОЛІНЕАРНІСТЬ РЕГРЕСОРІВ

При побудові регресійної моделі намагаються включити до неї найбільш впливові чинники, які інколи дублюють один одного, тобто, четверта умова 1МНК не виконується. Це означає, що або стовпчиковий ранг матриці $\boldsymbol{X}$ менше $\boldsymbol{N}$, або ж регресори сильно корелюють між собою. Якщо хоча б один стовпчик матриці $\boldsymbol{X}$ є лінійною комбінацією декількох, то має місце повна колінеарність. При цьому неможливо отримати оцінки коефіцієнтів моделі. Тоді необхідно уважно проаналізувати специфікацію моделі і зв'язки між чинниками регресії.

Повна колінеарність на практиці зустрічається досить рідко. Частіше зустрічається випадок, коли матриця $\boldsymbol{X}$ має повний стовпчиковий ранг, але між регресорами існує суттєва кореляція. У цьому випадку мова йде про мультиколінеарність регресорів. У цьому випадку формально можна розраховувати параметри моделі, але вони будуть мати «небажані» властивості.

Причини виникнення мультиколінеарності

Причинами мультиколінеарності можуть бути:

1) тенденція одночасної зміни економічних показників;
2) наявність трендів у динамічних рядах;
3) якщо декілька регресорів мають спільний часовий тренд, відносно якого вони здійснюють малі коливання;
4) використання в економетричних моделях лагових значень деяких чинників.

Наслідки мультиколінеарності

1. Оцінники дисперсій і коваріацій коефіцієнтів регресії досить великі.

Це означає, що будуть збільшуватися істинні і оцінники дисперсій і коваріацій коефіцієнтів регресії, а це призводить до таких наслідків: інтервали довіри коефіцієнтів регресії і прогнозні інтервали регресанда збільшуються, тестування стає неможливим, і це при великих значеннях $\boldsymbol{R^2}$.

2. Коефіцієнти регресії стають нестабільними.

Це означає, що невеликі зміни у специфікації моделі або у кількості спостережень призводять до великих змін значень коефіцієнтів регресії, до неправильних їхніх знаків і великих значень.

3. Труднощі в оцінці коефіцієнтів.

При повній колінеарності коефіцієнти регресії неможливо оцінити за методом 1МНК.

4. 1МНК-оцінники параметрів моделі втрачають властивість незміщеності.

5. Неможливо визначити внесок кожної незалежної змінної в дисперсію залежної змінної.

6. Оцінки параметрів при неколінеарних регресорах стають незначущими.

Ознаки мультиколінеарності

Розглянемо декілька рекомендацій для виявлення мультиколінеарності регресорів.

1. Аналізують матрицю парних коефіцієнтів кореляції незалежних змінних. Якщо значення коефіцієнта кореляції за абсолютною величиною перевищує 0.8, то існує мультиколінеарність.

2. Якщо визначник матриці $\boldsymbol{X^T X}$ близький до нуля, то це ознака мультиколінеарності.

3. Власні числа матриці $\boldsymbol{X^T X}$ відіграють важливу роль при визначенні мультиколінеарності. Якщо мінімальне власне число близьке до нуля, то це вказує на мультиколінеарність.

4. Велике значення коефіцієнта детермінації $\boldsymbol{R^2}$ і незначущість деяких коефіцієнтів регресії є ознакою мультиколінеарності.

5. Відомо, що детермінант кореляційної матриці регресорів задовольняє умові: $\boldsymbol{0 \le \|r\| \le 1}$. Якщо $\boldsymbol{\|r\|}$ близький до нуля, то це ознака мультиколінеарності регресорів.

6. Якщо модель двофакторна, то значення $\boldsymbol{r_{x_2 x_3}}$ достатнє для остаточного визначення мультиколінеарності.

7. Якщо коефіцієнт детермінації великий, а часткові коефіцієнти кореляції низькі, то мультиколінеарність можлива. Також можливо, що у моделі є надлишкові змінні. Але якщо коефіцієнт детермінації високий і часткові коефіцієнти кореляції високі, то мультиколінеарність не завжди можна виявити.

Метод інфляційних факторів визначення мультиколінеарності

1. Розглядається регресійне рівняння

$$\boldsymbol{y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_N x_N + u}.$$

Для чинників $\boldsymbol{x_2, \ldots, x_N}$ будуємо багатофакторні моделі

$$\boldsymbol{x_i = \gamma_1 x_1 + \ldots + \gamma_{i-1} x_{i-1} + \gamma_{i+1} x_{i+1} + \ldots + \gamma_N x_N + \varepsilon,\ i = \overline{2, N}}. \quad (*)$$

2. Методом 1МНК оцінюємо коефіцієнти кожної з моделей (*) і розраховуємо коефіцієнти детермінації $\boldsymbol{R_i^2}$, далі обчислюємо інфляційні фактори

$$\boldsymbol{iF_i = \frac{1}{1 - R_i^2}}.$$

Якщо $\boldsymbol{\min iF_i > 5\ \left(i = \overline{2, N}\right)}$, то вважаємо, що існує мультиколінеарність регресорів.

Звичайно зрозуміло, що при дослідженні якогось явища доцільніше користуватися статистичними тестами. Тому розглянемо тест Фаррара – Глобера.

Тест Фаррара – Глобера на наявність мультиколінеарності регресорів

Найбільш повне дослідження мультиколінеарності регресорів можна здійснити за допомогою тесту Фаррара – Глобера. Тест складається з трьох статистичних критеріїв: $\boldsymbol{\chi^2}$-критерій перевіряє весь масив регресорів на мультиколінеарність; $\boldsymbol{F}$-критерій – кожного регресора з масивом інших; $\boldsymbol{t}$-критерій – мультиколінеарність кожної пари регресорів.

Опишемо алгоритм тесту.

1. Нормалізація незалежних змінних.

Позначимо вектори незалежних змінних в економетричній моделі через $X_1, X_2, \ldots, X_N$. Компоненти k-го вектора X_k нормалізуються за формулою

$$x^*_{ki} = \frac{x_{ki} - \bar{x}_k}{\sigma_{x_k}},$$

де $\bar{x}_k$ – середнє арифметичне значення компонент k-ї незалежної змінної, $i = \overline{1,T}$, σ_{x_k} – середнє квадратичне відхилення k-ї незалежної змінної, T – кількість спостережень, а N – кількість незалежних змінних.

2. Обчислення кореляційної матриці.

Кореляційна матриця обчислюється за формулою

$$r = \frac{1}{T-1} X^{*T} X^*,$$

де X^* – матриця нормалізованих незалежних змінних.

3. Застосовуємо χ^2-критерій.

Розраховуємо χ^2_{cm} за формулою

$$\chi^2_{cm} = -\left[T - 1 - \frac{1}{6}\left(2N + 5\right)\right] \log \|r\|,$$

де $\|r\|$ – детермінант кореляційної матриці r. Якщо $\chi^2_{cm} > \chi^2_{kp} = \chi^2\left(\alpha;\, 0.5N\left(N-1\right)\right)$, то з ймовірністю $\left(1-\alpha\right)100\%$ у масиві незалежних змінних існує мультиколінеарність.

4. Визначаємо матрицю Z.

Матриця Z є оберненою до кореляційної матриці r: $Z = r^{-1}$.

5. Застосування F_k-критеріїв.

Використовуючи елементи матриці Z, розраховуємо F_k-статистики:

$$F_{kcm} = \left(z_{kk} - 1\right)\frac{T-N}{N-1}, \left(k = \overline{1,N}\right),$$

де z_{kk} – діагональні елементи матриці Z. Якщо $F_{kcm} > F_{kp} = F\left(\alpha; N-1; T-N\right)$, то k-та незалежна змінна мультиколінеарна з масивом інших змінних.

6. Дослідження за t-критерієм.

Розраховуємо часткові коефіцієнти кореляції:

$$r_{ki|\bullet} = \frac{-z_{ki}}{\sqrt{z_{kk} \cdot z_{ii}}}$$

і розраховуємо t_{ki}-статистики:

$$t_{kicm} = \left| r_{ki|\bullet} \right| \sqrt{\frac{T-N}{1-r_{ki|\bullet}^2}}.$$

Якщо $t_{kicm} > t_{kp} = t\left(\frac{\alpha}{2}; T-N\right)$, то коефіцієнт $r_{ki|\bullet}$ значимо відрізняється від нуля з ймовірністю $(1-\alpha)100\%$ і по величині $\left| r_{ki|\bullet} \right|$ будемо судити про наявність чи відсутність зв'язку між X_k і X_i.

Зауваження. Елементи кореляційної матриці r краще розраховувати за формулами теми 1. У цьому випадку елементи кореляційної матриці будуть розраховані більш точно.

Методи усунення мультиколінеарності

Що ж робити при наявності мультиколінеарності? Однозначної відповіді на це питання не існує. Представники деяких шкіл вважають, що нічого не потрібно робити, а наявність мультиколінеарності це суть нашого буття.

Але все ж таки потрібно обережно підходити до проблеми усунення мультиколінеарності. Якщо ми бажаємо усунути «зайві» змінні, то у деяких ситуаціях модель втратить свою економічну суть, або ж отримаємо модель зі зміщеними оцінками. Якщо за побудованою моделлю бажаємо робити тільки прогноз, то можна використовувати модель і при наявності мультиколінеарності. Тільки необхідно слідкувати, щоб залежність між мультиколінеарними чинниками для прогнозних значень зберігалася. Якщо коефіцієнт детермінації моделі великий, а коефіцієнти моделі значущі, то можна не звертати увагу на мультиколінеарність. Але якщо ми бажаємо аналізувати параметри моделі, то мультиколінеарність для нас уже проблема.

Розглянемо декілька рекомендацій по усуненню мультиколінеарності.

1. Змінити специфікацію моделі так, щоб знизити мультиколінеарність змінних до припустимої величини, тобто, щоб мультиколінеарність не була для нас проблемою.

2. Якщо між двома змінними X_i і X_j існує мультиколінеарність, то одну із них виключають з моделі. Яку змінну залишити у моделі, визначають економічною доцільністю. Якщо цього не можна зробити, то залишають ту, яка сильніше корелює з регресандом.

3. Інколи перетворення змінних знижує або і зовсім усуває мультиколінеарність, наприклад, від рівняння

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \ldots + \beta_N x_{Nt} + u_t,\ t = \overline{1,T}$$

переходять до рівняння перших різниць

$$y_t - y_{t-1} = \beta_2 \left(x_{2t} - x_{2,t-1} \right) + \ldots \beta_N \left(x_{Nt} - x_{N,t-1} \right) + u_t,\ t = \overline{1,T}.$$

4. Для зменшення мультиколінеарності можна перейти від незміщених оцінок, визначених методом 1МНК, до зміщених оцінок, що мають менше розсіювання відносно оціненого параметру.

Застосовують методи оцінки параметрів моделі, що враховують мультиколінеарність – «рідж-оцінки». При використанні «рідж-регресії» (або «гребневої регресії») замість незміщених оцінок розглядають зміщені оцінки, що задаються вектором

$\beta_\tau = \left(X^T X + \tau I\right)^{-1} X^T Y$, де τ – деяке додатне число, I – одинична матриця. Додача τ до діагональних елементів матриці $X^T X$ робить оцінки параметрів моделі зміщеними, але при цьому збільшується детермінант матриці системи нормальних рівнянь, з якої визначається параметр β – замість $\left\|X^T X\right\|$ він буде дорівнювати $\left\|X^T X + \tau I\right\|$.

5. Збільшення кількості спостережень може понизити або ж і усунути мультиколінеарність. Якщо ми використовуємо часові ряди, то можна зменшити довжину кожного періоду. Але тут необхідно бути обережним, тому що може з'явитися автокореляція.

6. При кореляції X_i і X_j перевіряють наявність мультиколінеарності між X_i і $X^* = X_i - X_j$. При присутності її виключають з моделі одну із змінних, в іншому разі замість X_j використовують змінну X^*.

7. Інколи можна перейти до сумісних рівнянь регресії, тобто до рівнянь зі взаємодією чинників, тобто до добутків чинників, якщо коефіцієнти при них будуть значимо відрізнятися від нуля.

8. Для усунення мультиколінеарності можна перейти до нових змінних, що ортогональні між собою і є лінійними комбінаціями початкових змінних. Це можна зробити за допомогою методу головних компонент.

Метод головних компонент дає можливість перейти від змінних X_1, X_2, …, X_N, що корелюють між собою, до лінійно незалежних змінних. Для цього, використовуючи змінні $X_1, X_2, \ldots, X_N$, побудуємо усі можливі нормовано-центровані лінійні комбінації.

Означення. Першою головною компонентою Z_1 системи показників $X_1, \ldots, X_N$ називається така нормовано-центрована лінійна комбінація цих показників, що має найбільшу дисперсію серед усіх нормовано-центрованих лінійних комбінацій.

Означення. k-ю головною компонентою Z_k системи показників $X_1, \ldots, X_N$ називається така нормовано-центрована лінійна комбінація цих показників, що не корелює з $k-1$ попередніми головними компонентами і серед усіх нормовано-центрованих лінійних комбінацій, що не корелюють з $k-1$ попередніми головними компонентами, має найбільшу дисперсію.

Алгоритм побудови головних компонент такий:

1) нормалізуємо змінні $X_1, \ldots, X_N$ за правилом:

$$x_{ij}^H = \frac{x_{ij} - \bar{x}_j}{\sigma_{x_j}}, \quad j = \overline{1, N},$$

отримаємо матрицю X^H;

2) будуємо кореляційну матрицю r:

$$r = \frac{1}{T-1}\left(X^H\right)^T X^H$$

або краще елементи (коефіцієнти кореляції між X_i та X_j) матриці розраховувати за формулами, описаними у темі 1;

3) знаходимо власні числа і власні вектори матриці r;
4) ранжуємо власні числа в порядку зменшення їх величин: $\lambda_1, \lambda_2, \ldots, \lambda_N$;
5) будуємо головні компоненти

$$Z_k = X^T l_k = x_1 l_{k1} + x_2 l_{k2} + \ldots + x_N l_{kN},$$

де l_k – власний вектор матриці r, що відповідає власному числу λ_k.

Зауважимо, що дисперсія k-ї головної компоненти Z_k дорівнює λ_k.

Головні компоненти характеризуються такими властивостями:

1) їх кількість дорівнює кількості вихідних ознак;
2) вони є ортогональними;
3) вони нормалізовані (середні значення рівні нулеві, дисперсії дорівнюють одиниці);
4) вони впорядковані таким чином, що перша головна компонента пояснює найбільшу частку дисперсії вихідних ознак, наступна – найбільшу частку дисперсії, що залишилася непоясненою першою компонентою, і т. д.;
5) якщо початкові дані нормалізовані, то коваріаційні і кореляційні матриці співпадають і $\sum_{k=1}^{N} \lambda_k = N$.

Відомо також, що $\sum_{i=1}^{N} \sigma_x^2 = \sum_{i=1}^{N} \sigma_z^2 = \sum_{k=1}^{N} \lambda_k = N$.

Враховуючи зменшення частки пояснюваної дисперсії вихідних ознак наступною головною компонентою, на практиці для аналізу беруться не всі компоненти, а лише ті, що пояснюють наперед задану сумарну частку цієї дисперсії. Ми користуємося критерієм інформативності:

$$I_p = \frac{\lambda_1 + \ldots + \lambda_p}{N}.$$

На практиці прийнято, що якщо $I_p \geq 0.8$, то кількості головних компонент достатньо, щоб вони пояснювали дисперсію змінних не менше, ніж на 80 %.

Розглянемо матрицю навантажень (факторних навантажень) $A = \left(a_{ij}\right),\ i, j = \overline{1, N}$ головних компонент на початкові ознаки, яка є важливою характеристикою головних компонент. Якщо головні компоненти будуються для нормалізованих ознак, то елементи матриці навантажень a_{ij} визначають ступінь тісноти парного лінійного зв'язку між X_i і Z_j.

Матриця навантажень A визначається співвідношенням $A = L^T \Lambda^{\frac{1}{2}}$, де матриця L складається з рядків $l_j,\ j = \overline{1, N}$, що є власними векторами матриці r, а

$$\Lambda^{\frac{1}{2}} = \begin{pmatrix} \sqrt{\lambda_1} & 0 & \dots & 0 \\ 0 & \sqrt{\lambda_2} & \dots & 0 \\ \dots & \dots & \ddots & \dots \\ 0 & 0 & \dots & \sqrt{\lambda_N} \end{pmatrix}.$$

Маємо

$$A = \begin{array}{c} \\ X_1 \\ X_2 \\ \vdots \\ X_N \end{array} \begin{array}{c} \begin{array}{cccc} Z_1 & Z_2 & \dots & Z_N \end{array} \\ \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1N} \\ a_{21} & a_{22} & \dots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \dots & a_{NN} \end{pmatrix} \end{array}$$

Так як головні компоненти є лінійною комбінацією ознак, то досить важко дати економічне тлумачення головних компонент. Щоб все ж таки це зробити, виділяємо ті ознаки X_i, які досить сильно корелюють ($r_{X_i Z_j} \geq 0.6$) з головною компонентою Z_j, і по назвам ознак намагаємося дати узагальнену назву головній компоненті Z_j.

Приклад 12. Використовуючи метод Фаррара – Глобера, дослідити вектори x_2, x_3, x_4, x_5 (див. приклад 1) на мультиколінеарність, і якщо вона існує, то усунути її за допомогою методу головних компонент.

Згідно з алгоритмом Фаррара – Глобера спочатку будуємо кореляційну матрицю

$$r = \begin{pmatrix} 1.0000 & -0.5228 & -0.2140 & 0.0993 \\ -0.5228 & 1.0000 & 0.2275 & 0.0131 \\ -0.2140 & 0.2275 & 1.0000 & 0.0141 \\ 0.0993 & 0.0131 & 0.0141 & 1.0000 \end{pmatrix}.$$

Для визначення мультиколінеарності у масиві змінних розрахуємо χ^2_{cm}: $\chi^2_{cm} = 8.7792$. Далі, $\chi^2_{kp} = \chi^2\left(0.05; 6\right) = 1.64$. Так як $\chi^2_{cm} > \chi^2_{kp}$, то з ймовірністю 95 % у масиві змінних існує мультиколінеарність.

Далі розрахуємо матрицю $Z = r^{-1}$:

$$Z = \begin{pmatrix} 1.4171 & 0.7102 & 0.1438 & -0.1521 \\ 0.7102 & 1.4106 & -0.1678 & -0.0867 \\ 0.1438 & -0.1678 & 1.0693 & -0.0272 \\ -0.1521 & -0.0867 & -0.0272 & 1.0166 \end{pmatrix}.$$

Розглянемо питання про можливу мультиколінеарність між x_i ($i = \overline{2,5}$) і масивом інших змінних. Для цього розрахуємо F_{icm}: $F_{2cm} = 2.6419$, $F_{3cm} = 2.6003$, $F_{4cm} = 0.4391$, $F_{5cm} = 0.1053$, а $F_{kp} = 2.85$. Так як F_{icm} ($i = \overline{2,5}$) $< F_{kp}$, то жодна зі змінних не мультикорелює з відповідним масивом інших змінних.

Знайдемо часткові коефіцієнти кореляції: $r_{x_2x_3|\bullet} = -0.5023$, $r_{x_2x_4|\bullet} = -0.1168$, $r_{x_2x_5|\bullet} = 0.1267$, $r_{x_3x_4|\bullet} = 0.1366$, $r_{x_3x_5|\bullet} = 0.0724$, $r_{x_4x_5|\bullet} = 0.0261$. Далі розраховуємо t_{cm} для всіх $r_{x_ix_j|\bullet}$. Отримаємо: $t_{cmx_2x_3} = 2.5320$, $t_{cmx_2x_4} = 0.5127$, $t_{cmx_2x_5} = 0.5567$, $t_{cmx_3x_4} = 0.6010$, $t_{cmx_3x_5} = 0.3163$, $t_{cmx_4x_5} = 0.1138$, а $t_{kp} = t\left(0.05; 21\right) = 1.721$, тобто з ймовірністю 95 % коефіцієнт $r_{x_2x_3|\bullet}$ значимо відрізняється від нуля.

Отже, два з трьох критеріїв вказують на те, що існує мультиколінеарність векторів. Усувати її будемо методом головних компонент. Для цього розрахуємо власні числа і власні вектори матриці r. Проранжувавши власні числа, отримаємо:

$$\lambda_1 = 1.6722,\ l_1 = \left(0.6411, -0.6389, -0.4188, 0.0734\right)^T,$$

$$\lambda_2 = 1.0178,\ l_2 = \left(0.1268, 0.0786, 0.2422, 0.9587\right)^T,$$

$$\lambda_3 = 0.8448,\ l_3 = \left(-0.2803, 0.3190, -0.8751, 0.2320\right)^T,$$

$$\lambda_4 = 0.4653,\ l_4 = \left(-0.7031, -0.6956, 0.0107, 0.1473\right)^T.$$

Тоді головні компоненти матимуть вигляд:

$$z_1 = 0.6411\,x_2 - 0.6389\,x_3 - 0.4188\,x_4 + 0.0734\,x_5,$$

$$z_2 = 0.1268\,x_2 + 0.0786\,x_3 + 0.2422\,x_4 + 0.9587\,x_5,$$

$$z_3 = -0.2803\,x_2 + 0.3190\,x_3 - 0.8751\,x_4 + 0.2320\,x_5,$$

$$z_4 = -0.7031\,x_2 - 0.6956\,x_3 + 0.0107\,x_4 + 0.1473\,x_5.$$

Використовуючи критерій інформативності, вияснимо питання про те, скільки головних компонент достатньо взяти, щоб вони найбільш інформативно пояснювали дисперсію змінних z_1, z_2, z_3, z_4 (однаково, що і x_2, x_3, x_4, x_5). Практика показує, що якщо p перших головних компонент пояснюють на 80 % і більше всю дисперсію головних

компонент, то достатньо цих p головних компонент для подальшого дослідження об'єкта. Для цього розрахуємо критерій інформативності $I_p = \frac{\lambda_1 + \lambda_2 + ... + \lambda_p}{N}$. Отримали, що $I_3 = \frac{1.6722 + 1.0178 + 0.8448}{4} = 0.8837 > 0.8$. Отже, три перші головні компоненти на 88.37 % пояснюють всю дисперсію головних компонент.

Далі розглянемо матрицю навантажень A головних компонент на початкові ознаки: $A = L^T \Lambda^{\frac{1}{2}}$. Виписуємо результати тільки для 3 перших головних компонент.

$$A = \begin{array}{c} \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array} \begin{array}{c} \begin{array}{ccc} Z_1 & Z_2 & Z_3 \end{array} \\ \begin{pmatrix} 0.8290 & 0.0740 & -0.3849 \\ -0.9092 & 0.1486 & 0.0098 \\ -0.3625 & 0.2341 & -0.8043 \\ 0.1640 & 0.9672 & 0.2226 \end{pmatrix} \end{array}.$$

Отримали, що перша головна компонента досить сильно корелює зі змінними x_2, x_3. Друга головна компонента – з x_5. Третя – з x_4. Тепер досить легко дати економічне тлумачення першим трьом головним компонентам.

РЕКОМЕНДОВАНА ЛІТЕРАТУРА

1. Грубер Й. Эконометрия. Введение в эконометрию. Т. 1. – К., 1996. – 397 с.
2. Доугерти К. Введение в эконометрику. – М.: МГУ, 1999. – 102 с.
3. Под ред. Н. Н. Елисеевой. Эконометрика. – М.: Финансы и статистика, 2002. – 242 с.
4. Кремер Н. Ш., Путко Б. А. Эконометрика. – М.: ЮНИТИ, 2002. – 311 с.
5. Лук'яненко І., Краснікова Л. Економетрика. – К.: Знання, 1998. – 494 с.
6. Магнус Я. Р., Катышев П. К., Перестецкий А. А. Эконометрика. Начальный курс. – М.: Дело, 1998. – 246 с.
7. Толбатов Ю. А. Економетрика. – К.: Четверта хвиля, 1997. – 219 с.

ЗМІСТ