Көптік регрессия. Факторларды іріктеп сұрыптау

Көптік регрессияда, бір қорытынды *y* факторы мен бірнеше $x_1, x_2,\ldots,x_n$ тәуелсіз факторлардың арасындағы байланыс анықталады, яғни $y=f(x_1, x_2,\ldots,x_n)$ түріндегі байланыс анықталуы керек. $y=f(x_1, x_2,\ldots,x_n)$ - көптік регрессия теңдеуі.

***n* рет байқау жүргізіліп, соның нәтиежесінде төмендегідей мағлуматтар алынған болсын делік:**

n байқау	*Y* қорытынды фактор	Тәуелсіз факторлар			
1 *2* *...* *N*				... ...	

Көптік регрессия моделін құрғанда осы тәуелсіз факторлардың бәрі қажет пе, жоқ па деген сұрақ туады. Зерттеуші факторлардың экономикалық мағыналарына қарап олардың арасындағы байланыстар туралы тұжырым жасай алады. Ол қорытынды фактор мен тәуелсіз факторлардың қайсысы керек, қайсысы керегі жоқ, қарастырылып отырған тәуелсіз факторлар жеткілікті ме, жоқ па, солар туралы қорытынды жасауы үшін жеткілікті білімі мен тәжірибесі болуы керек

Артық факторлар алынып тасталуы керек (олар регрессия теңдеуінің қате болжам жасауына соқтырады). Егер факторлар жеткіліксіз болатын болса жаңадан байқаулар жасап қосымша мағлұматтар жиналуы керек.

Факторлады іріктеп, сұрыптаған кезде төмендегідей талаптар қанағаттандырылатын болуы керек:

1. Факторлардың мәндері сандармен өрнектелетін болуы керек. Байқаулардан алынған, санмен өрнектелмей тұрған сапалық белгілерді сандық белгілеулерге айналдыру керек.

2. Тәуелсіз фактор қорытынды фактормен тығыз байлаысты болуы керек.

3. Тәуелсіз факторлардың арасында өзара тығыз байланыс болмауы керек. Егер модельде арасындағы байланыс тығыз екі фактор бар болса олардың бірі модельден шығарылуы керек.

4. Тәуелсіз факторлар арсында мультиколлинеарлық байланыс болмауы керек (мультиколлинеарлық ұғымын кейінірек анықтаймыз).

Факторларды іріктеу жұмысы екі этаптан тұрады. Бірінші этапта экономикалық мағынасына қарап факторлар іріктеп сұрыпталады. Жұмысты зерттеуші өзінің білімі мен тәжірибесіне сүйеніп жасайды

Екінші этапта математикалық аппарат қолданылады. Ол үшін жұптық сызықтық корреляция матрицасы құрылады (ол матрица қалай құрылатынын алдыңғы тақырыптарда айтып кеткен болатынбыз). Осы матрицаның көмегі арқылы тәуелсіз факторлардың қос-қостан арасындағы байланыстарының тығыздығы анықталады. Егер олардың арасындағы байланыс тығыз болса *($|r(x_i,x_j)|\geq 0,7$)*, онда бұндай айнымалыларды (факторларды) өзара тәуелсіз факторлардеп есептеуге болады (оларды коллинеар факторлар деп атайды).

Жалпы *0,7* саны туралы зерттеушілер арасында толық келісім жоқ, кейде оның орнына *0,6,* кейде *0,8* шамасын қарастыруға болады. Әрбір дербес жағдайда айнымалылардың экономикалық мағынасы ескерілуі керек.

Корреляциялық матрицаның фо

$$R=\begin{pmatrix} 1 & r_{yx_1} & r_{yx_2} & \dots & r_{yx_n} \\ r_{yx_1} & 1 & r_{x_1x_2} & \dots & r_{x_1x_n} \\ r_{yx_2} & r_{x_1x_2} & 1 & \dots & r_{x_2x_n} \\ \dots & \dots & \dots & \dots & \dots \\ r_{yx_n} & r_{x_1x_n} & r_{x_2x_n} & \dots & 1 \end{pmatrix}$$

Корреляциялық матрица диагоналына қарағанда симметриялы, сондықтан оның диагоналының үстіндегі, немесе астындағы элементтерді жазбайды.

Корреляциялық матрицаны MS Excel программасын қолданып тез есептеуге болады.

Егер факторлар коллинеар болатын болса, олар модельде бірдей қызмет атқарады деп есептеуге болады. Сондықтан олардың біреуі ғана модельде қалып, қалғаны модельден шығарылуы керек. Әрине бұндай кезде қайсысы шығарылуы керек деге сұрақ туады. Қалатын тәуелсіз фактормен қорытынды фактордың арасындағы байланыс тығызырақ болуы керек (басқасымен салыстырғанда). Сосын қалған факторлармен арадағы байланысының тығыздығы аз болуы керек (басқасымен салыстырғанда).

Айталық x_i, x_j факторының коллинеар болсын $(|r(x_i, x_j)| \geq 0{,}7)$. Онда мынадай байланыстар тексеріледі:

r_{yxi}, r_{yxj} осылардың қайсысы үлкен (қайсысы тәуелді айнымалымен тығызырақ байланыста), сол қалуы керек.

r_{x1xi}, r_{x2xi},…,r_{xnxi} (ішінде r_{xixj} жоқ), r_{x1xj}, r_{x2xj},…,r_{xnxj} (ішінде r_{xixj} жоқ). Осылардың абсолют шамасы қалатын фактор үшін мейлінше аз болуы керек.

Яғни факторды модельден шығарғанда осы екі талаптың екеуін де қанағаттандыруға тырысу керек.

Мысал. 4 фактор берілген. Олардың жұптық сызықтық корреляция коэффициенттерінің матрицасы мына түрде берілген

	1				
	0,75	1			
	0,72	0,3	1		
	0,79	0,71	0,81	1	
	0,58	0,12	0,07	0,13	1

Матрицаға қарасақ x_3 факторы x_1 мен x_2 факторларымен тығыз байланысты. $r_{x3x1}=0,71$; $r_{x3x2}=0,81$. Яғни x_1 мен x_2 өзгерген кезде олармен тығыз байланыста тұрған x_3 – те өзгереді.

Енді осы үшеуінің қайсысы модельден шығуы керек. Ең қорытынды фактормен тығыз байланысқаны x_3. Бірақ соған қарамай ол модельден шығарылады. Қараңыз: $r_{x3x2}>r_{x2y}$ яғни x_3 тің x_2 мен байланысы қорытынды фактормен байланысынан тығызырақ.

Сонымен жұптық сызықтық корреляция матрицасының көмегі арқылы факторларды сұрыптаған кезде қай фактордың регрессия моделінен шығуы керек екенін анықтауға болады екен.

Мультиколлинеарлық байланыс

Кей жағдайда жалпы сандары екеуден көп болатын тәуелсіз факторлар жиынтық құрайды да, сол жиынтықтың ішіндегі біреуіне қалғандары жабылып мықты әсер етеді, яғни олардың арасында бірден көзге көріне қоймайтын тәуелділік бар. Мұндай тәуелділікті мультиколлинеарлық тәуелділік дейді.

Мультиколлинеарлық тәуелділік болған кезде факторлар топтасып, қорытынды факторға әсер етеді. Мұның төмендегідей салдары бар:

- регрессия параметрлерінің баламалары сенімсіз баламалар болады;
- әрбір тәуелсіз фактордың қорытынды факторға қаншалықты әсер ететінін бағалау мүмкін болмай қалады.

Іс жүзінде мультиколлинеарлық тәуелділік бар жоғын білу үшін тәуелсіз факторлардың жұптық сызықтық корреляция коэффициенттерінің матрицасын зерттейді. Егер дәлірек айтатын болсақ корреляциялық матрицаның анықтауышын зерттейді.

Қарастырып отырған көптік регрессия моделі төмендегідей болсын:

$$y=a+b_1x_1+b_2x_2+b_3x_3+\varepsilon$$

Осы модель үшін корреляция коэффициенттерінің матрицасының анықтауышын есептелік:

$$det|R| = \begin{vmatrix} r_{x_1x_1} & r_{x_1x_2} & r_{x_1x_3} \\ r_{x_2x_1} & r_{x_2x_2} & r_{x_2x_3} \\ r_{x_3x_1} & r_{x_3x_2} & r_{x_3x_3} \end{vmatrix} = \begin{vmatrix} 1 & r_{x_1x_2} & r_{x_1x_3} \\ r_{x_2x_1} & 1 & r_{x_2x_3} \\ r_{x_3x_1} & r_{x_3x_2} & 1 \end{vmatrix}$$

Егер факторлар өзара тәуелсіз болса, онда

$$r_{x_ix_j} = 0, \quad i \neq j \qquad \Rightarrow \qquad det|R| = \begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} = 1$$

Енді факторлар бірімен бірі толық тәуелді болсын делік, онда

$$r_{x_ix_j} = 1, i \neq j \qquad \Rightarrow \qquad det|R| = \begin{vmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{vmatrix} = 0$$

Сонымен осы мысалдан мынандай қорытынды жасауға болады екен: $det|R|$ шамасы нольге таяу болса, онда тәуелсіз айнымалылар арасында мультиколлинеарлық байланыс тым аз болғаны. Егер $det|R|$ шамасы бір санына таяу болса, онда олардың арасында мультиколлинеарлық байланыс тығызырақ болғаны.

Егер мультиколлинеарлық байланыс бар болса, онда модельге тек мультиколлинеарлыққа аз ғана әсер етіп тұрған факторларды ғана енгізу керек. Мұнда мультиколлинеарлықты өсіріп тұрған фактордың басқа факторлармен арасындағы сызықтық корреляция коэффициентінің абсолют шамасы басқаларына қарағанда үлкен болатыны ескерілуі керек.

Бұрын қарастырылған мысалдағы корреляциялық матрицаның анықтауышын есептелік:

$$det|R| = \begin{vmatrix} 1 & 0,3 & 0,71 & 0,12 \\ 0.3 & 1 & 0,81 & 0,07 \\ 0,71 & 0,81 & 1 & 0,13 \\ 0,12 & 0,07 & 0,13 & 1 \end{vmatrix} = 0,093$$

Бұл анықтауыштың абсолют шамасы нольге жуық болғандықтан қарастырып отырған тәуелсіз факторлардың арасында мультиколлинеарлық байланыс жоқ деп есептеуге болады.