УДК 519.23
DOI: 10.30838/J.BPSACEA.2312.261119.78.592

МОДЕЛЬ МНОЖИННОЇ РЕГРЕСІЇ НА КАТЕГОРІЙНІ ФАКТОРИ

ЦИБРІЙ Л. В., *к. ф.-м. н., доц.*

Кафедра прикладної математики та інформаційних технологій, Державний вищий навчальний заклад «Придніпровська державна академія будівництва та архітектури», вул. Чернишевського, 24-а, 49600, Дніпро, Україна, тел. +38 (056) 756-34-10, e-mail: tsybrii.larisa@pgasa.dp.ua, ORCID ID: 0000-0002-7427-0770

Анотація. ***Мета дослідження.*** Знайти кількісну оцінку впливу категорійних факторів на пояснювану змінну. Для передбачення значення числової змінної залежно від значень інших числових змінних використовується регресійний аналіз статистичної моделі, побудованої за даними спостережень. Однак у багатьох ситуаціях у модель необхідно включати категорійні змінні. ***Метод.*** Категорійні фактори включаються в модель як фіктивні змінні, значення яких 1 або 0 відповідають наявності або відсутності певного категоріального рівня пояснювальної змінної. Однак такий підхід виправданий лише в тому випадку, коли у категорійної змінної тільки два рівні: наявність або відсутність певної якості. Тоді значення фіктивної змінної повинне дорівнювати 1 у першому випадку й 0 – у другому. За наявності декількох категоріальних рівнів пропонується в модель множинної регресії вводити фіктивні змінні, відповідні кожному рівню і які набирають значення 1, якщо категорійна пояснювальна змінна набирає значення відповідного рівня, і дорівнюють 0 для всіх інших її значень (рівнів). Крім того, модель дозволяє врахувати можливу взаємодію категоріальних рівнів різних пояснювальних нечислових змінних і ефект такої взаємодії за кількісної оцінки їх впливу на пояснювану змінну. ***Результати.*** Запропоновано модель множинної регресії на категорійні фактори, що дозволяє знайти кількісну оцінку впливу на пояснювану змінну не тільки числових, а й категорійних незалежних змінних. Причому модель враховує той факт, що за відсутності всякого впливу з боку пояснювальних змінних, тобто коли вони всі рівні 0, регресія також повинна дорівнювати 0. При цьому так званий «зсув» прямої регресії відсутній (і тим більше не може бути від'ємним, як для математичної прямої). ***Наукова новизна.*** Пропонована модель розширює й узагальнює можливості регресійного аналізу статистичних моделей на випадок категорійних факторів. ***Практична значимість.*** Модель множинної регресії на категорійні фактори дозволяє вирішувати питання вибору категоріальних рівнів нечислових параметрів для проектування систем; ухвалення рішення про вибір стратегії у фінансовій діяльності й в управлінні.

Ключові слова: *множинна регресія; категорійний фактор; категоріальні рівні; пояснювана змінна; пояснювальні змінні; кількісна оцінка*

MULTIPLE REGRESSION MODEL ON CATEGORICAL FACTORS

TSYBRII L.V., *Cand. Sc. (Techg.), Ass. Prof.*

Department of Applied Mathematics and Information Technology. State Higher Educational Institution “Prydniprovska State Academy of Civil Engineering and Architecture”, 24-a, Chernyshevskoho St., 49600, Dnipro, Ukraine, tel. +38 (056) 756-34-10, e-mail: tsybrii.larisa@pgasa.dp.ua, ORCID ID: 0000-0002-7427-0770

Abstract. ***Purpose.*** To find a quantitative assessment of the influence of categorical factors on the variable being explained. To predict the value of a numerical variable depending on the values of other numerical variables, a regression analysis of a statistical model based on observational data is used. However, in many situations, categorical variables must be included in the model. ***Method.*** Categorical factors are included in the model as dummy variables whose values 1 or 0 correspond to the presence or absence of a certain categorical level of the explanatory variable.However, such an approach is justified only if the category variable has only two levels: the presence or absence of a certain quality. Then the value of the dummy variable is due to be equal to 1 in the first case and 0 − in the second one. If there are several categorical levels, it is proposed to introduce dummy variables corresponding to each level and taking the value 1 into the multiple regression model if the categorical explanatory variable takes the value of the corresponding level and is 0 for all its other values (levels). In addition, the model makes it possible to take into account the possible interaction of categorical levels of various explanatory non-numerical variables and the effect of such interaction when quantifying their influence on the explained variable. ***Results.*** Proposed model of multiple regression on the categorical factors , which allows to find a quantitative assessment of the impact on the explanatory variable, not only numerical but also categorical independent variables. Moreover, the model takes into account the fact that in the absence of any influence from the explanatory variables, i.e. when they are all 0, the regression should also be 0 . Moreover, the so-called “shift” of direct regression is absent (and, moreover, cannot be negative, as for the mathematical line) . ***Scientific novelty.*** The proposed model extends and generalizes the capabilities of the regression

analysis of statistical models for the case of categorical factors. ***Practical relevance.*** The model of multiple regression on categorical factors allows us to solve the following problems: selection of categorical levels of non-numeric parameters when designing systems; deciding on the choice of strategy in financial activities and management.

Keywords: *multiple regression; categorical factor ; categorical levels ; explained variable; explanatory variables; quantitative assessment*

Постановка проблеми. Модель множинної регресії дозволяє виконати найважливіше завдання моделювання – знайти оцінки показників функціонування системи. Наявність моделі дає можливість порушувати питання про управління системою хоча б у межах вибірки, тобто зафіксованих спостережень.

Модель множинної регресії тим точніша, чим більше пояснювальних (незалежних) змінних використовується для передбачення значення змінної, що пояснюється (залежної). Для створення моделі відбір числових змінних, що включаються у модель, виконується методами статистичного аналізу [2; 3]. За необхідності включити в модель категорійні фактори виникає проблема, як відбирати змінні, що набирають категоріальні значення, і як оцінити їх вплив на числову залежну змінну. Для вирішення першої частини проблеми можна використати дисперсійний аналіз, щоб виявити статистично значимі різниці між рівнями факторів або ефект взаємодії факторів.

Аналіз публікацій. В основі створення моделі системи лежить аналіз даних спостережень або прогону імітаційної моделі методами математичної статистики. Дисперсійний аналіз ANOVA дає можливість оцінити ситуацію із залежністю пояснюваної змінної від категорійних факторів у цілому, без кількісного порівняння впливу їх на стан системи. Додаткові дослідження за допомогою процедури Тьюкі – Крамера дозволяють попарно порівнювати вплив рівнів факторів тільки з погляду їх схожості або розходження, але не дозволяють зробити кількісне порівняння

Для передбачення значення числової залежної змінної за значеннями категорійної змінної необхідно замінити її категоріальні рівні числами. Звичайно вводиться одна фіктивна змінна, котра може набувають тільки одного із двох значень: 1 або 0 [3]. Такий підхід зручний тільки в тому випадку, коли рівнями категорійної змінної постає наявність або відсутність якої-небудь властивості, тобто коли можливими значеннями категорійної змінної є «так» або «ні».

Мета дослідження знайти кількісну оцінку впливу категорійних змінних на пояснювану числову змінну

Виклад основного матеріалу. У загальному випадку стан системи визначається значенням пояснюваної змінної й значеннями числових і категорійних пояснювальних змінних. Категорійні фактори звичайно мають два й більше рівнів, що характеризують певні властивості. Наприклад, великий, середній або малий капітал.

У таких ситуаціях ефект впливу кожної властивості можна врахувати, тільки якщо кожному рівню поставити у відповідність фіктивну змінну. Така змінна набирає значення $X = 1$ за наявності цієї властивості, тобто для одного конкретного рівня, і $X = 0$ в інших випадках, тобто для інших рівнів. Так, для прикладу з капіталом варто ввести 3 фіктивні змінні $X1$, $X2$ і $X3$: для великого капіталу $X1 = 1$, $X2 = 0$, $X3 = 0$; для середнього капіталу $X1 = 0$, $X2 = 1$, $X3 = 0$; для малого капіталу $X1 = 0$, $X2 = 0$, $X3 = 1$. При цьому $X1 + X2 + X3 = 1$, тобто тільки одна фіктивна змінна може дорівнювати 1, а інші одночасно із цим рівні 0.

У загальному випадку лінійна регресія є функцією k числових змінних $X_1, X_2, \ldots X_k$:

$$Y = M(Y/X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k . \quad (1)$$

Тут β_i ($i = 1, 2, \ldots, k$) – коефіцієнт чистої регресії, що являє собою зміну середнього значення змінної Y при зміні значення числової змінної X_i на одиницю її виміру при постійних інших пояснювальних змінних.

Для оцінки невідомих параметрів β_i ($i = 1, 2, \ldots, k$) рівняння регресії

використовується випадкова вибірка обсягом *n*. Модель множинної регресії для *m*-го (*m* = 1, 2, ..., *n*) спостереження можна подати так:

$$y_m = \beta_0 + \beta_1 x_{m1} + \beta_2 x_{m2} + ... + \beta_k x_{mk} + \varepsilon_m, \qquad (2)$$

де:

ε_m – випадкова помилка змінної *Y* в m-му спостереженні.

Для передбачення значень пояснюваної змінної залежно від категоріального значення (рівня) нечислової пояснювальної змінної використовується модель множинної регресії з фіктивними змінними. За наявності двох категорійних пояснювальних змінних двофакторний дисперсійний аналіз дозволяє перевірити значимість ефектів впливу кожної з них і їх взаємодії.

Якщо кожна з категорійних змінних має не менше двох рівнів, для визначення параметрів системи необхідно кожному рівню поставити у відповідність фіктивну змінну. Модель множинної регресії на числові й багаторівневі категорійні змінні може бути записана так:

$$y_m = \sum_{i=1}^{k} \beta_i X_i + \sum_{i=1}^{r} \beta_i^A X_i^A + \sum_{j=1}^{c} \beta_j^B X_j^B + \sum_{i=1}^{r} \sum_{j=1}^{c} \beta_{ij} X_i^A X_j^B + \varepsilon_m \quad (m = 1, 2, \ldots, n), \qquad (3)$$

де:

$X_i^A (i = 1, 2, ..., r)$ – фіктивні змінні, відповідні *i*-му рівню фактора А;

$X_j^B (j = 1, 2, ..., c)$ – фіктивні змінні, відповідні *j*-му рівню фактора В;

$X_i^A X_j^B$ – добуток фіктивних змінних, враховує взаємодію факторів А і В на всіх рівнях і набирає одне із двох значень: 0 або 1. Причому:

$$\sum_{i=1}^{r} X_i^A = 1 \; ; \; \sum_{j=1}^{c} X_j^B = 1 \; ; \; \sum_{i=1}^{r} \sum_{j=1}^{c} X_i^A X_j^B = 1. \qquad (4)$$

Де: β_i (i = 1,2,…,k) виконують ту ж роль, що в рівнянні (1); β_i^A; β_i^B *і* β_{ij} – коефіцієнти чистої регресії, що являють собою відповідно зміну значення змінної *Y* при *i*-му рівні фактора *A* ($X_i^A = 1$); при *j*-му рівні фактора *B* ($X_j^B = 1$) і при *i*-му рівні фактора *A* ($X_i^A = 1$) та *j*-му рівні фактора *B* ($X_j^B = 1$), тобто $X_i^A X_j^B = 1$; ε_m – випадкова помилка *m*-го спостереження.

Формула (3) не містить вільного члена й у випадку рівності всіх пояснювальних змінних нулю, тобто в ситуації, що відповідає відсутності самої системи, забезпечує значення пояснюваної змінної рівним нулю, що цілком логічно.

Оцінка невідомих коефіцієнтів β дасть розв'язання системи (3) методом найменших квадратів або за допомогою інструмента Регресія надбудови Excel Пакет аналізу. Перевірка їх значимості й значимості функції регресії виконується за допомогою *t*-критерію й критерію *F*, як звичайно для регресії на числові фактори.

Проаналізуємо роботу взаємних фондів, що володіють портфелем цінних паперів. Придбавши акції (частку) такого фонду, інвестор вступає у володіння всіма акціями компаній, що належать фонду. Необхідно вибрати фонд, у який варто вкладати кошти. Для цього спочатку необхідно порівняти ефективність взаємних фондів різних категорій: тих, що спеціалізуються на акціях великих, середніх або малих компаній; орієнтованих на швидкий або повільний ріст капіталу.

Проведемо аналіз річних показників прибутковості 105 фондів США в 2001 році, що характеризується дуже сильними коливаннями прибутковості взаємних фондів [3]. Розглянемо такі показники: ВИД (вид акцій, що належать фонду: великі, середні й малі компанії); МЕТА (мета фонду: швидкий або повільний ріст капіталу); АКТИВИ (млн доларів); ВИТРАТИ (витрати, понесені фондом, у % від середнього обсягу чистих активів); ПРИБУТКОВІСТЬ (у % за 12 місяців 2001 року). Уявлення про інформацію дасть фрагмент вихідних даних:

Фонд	Вид	Мета	Активи	Витрати	Прибутковість
ABN Amro Montag & Calswell Growth	великий	Швидкий	1184,9	0,77	-13,10
Aim Small Cap Growth A	малий	Швидкий	602,2	1,13	-13,80
Alger Large Cap Growth B	великий	Швидкий	614,0	1,96	-12,90
Amer. Century GiftTrust Inv.	малий	Швидкий	921,1	1,00	-35,4
Armada Large Cap value I	великий	Повільний	679,3	0,97	-3,8
AXP Equity Value A	великий	Повільний	1132,0	0,95	-4,4

Дисперсійний аналіз приводить до висновку: на прибутковість фонду впливають вид капіталу (фактор А), темп росту капіталу (фактор В) і їх взаємодія. Вводимо фіктивні змінні: X1, X2 і X3 , що набирають значення 1 для великого, середнього й малого капіталу відповідно; і *X*4 і *X*5, що набирають значення 1 для швидкого й повільного росту капіталу відповідно, і всі рівні 0 в інших випадках.

Активи	Витр.	X1	X2	X3	X4	X5	X1X4	X2X4	X3X4	X1X5	X2X5	X3X5	Прибутк.
1184,9	0,77	1	0	0	1	0	1	0	0	0	0	0	-13,10
602,2	1,13	0	0	1	1	0	0	0	1	0	0	0	-13,80
614,0	1,96	1	0	0	1	0	1	0	0	0	0	0	-12,90
921,1	1,00	0	0	1	1	0	0	0	1	0	0	0	-35,4
679,3	0,97	1	0	0	0	1	0	0	0	1	0	0	-3,8
1132,0	0,95	1	0	0	0	1	0	0	0	1	0	0	-4,4

Такий вигляд матиме матриця системи (3) для наведеного фрагмента вихідних даних. Розв'язавши систему методом найменших квадратів, дістаємо оцінку функції регресії прибутковості фонду на числові змінні Активи ($X_А$) і Витрати ($X_И$) і фіктивні змінні X1, X2 і X3 та X4 і X5, що відповідають категоріальним рівням факторів Вид і Мета:

$$Y = -0{,}00034X_А + 6{,}53385X_И - 7{,}68208X_1 - 5{,}33059X_2 + 0{,}72684X_3 - 15{,}108X_4 + 2{,}82207X_5 -$$
$$- 1{,}97871X_1X_4 - 10{,}022X_2X_4 - 3{,}10716X_3X_4 - 5{,}70334X_1X_5 + 4{,}60141X_2X_5 + 3{,}834X_3X_5. \quad (5)$$

Коефіцієнт детермінації $R^2 = 0{,}589$, що означає, що 58,9 % варіації прибутковості пояснюється змінами всього набору пояснювальних змінних. Внесок кожної з них оцінюється коефіцієнтом у рівнянні (5). Активи незначно впливають на прибутковість. Збільшення витрат на 1 % від середнього обсягу чистих активів виникає збільшення прибутковості на 6,53 %. Фонди, що володіють акціями великих компаній (*X*1=1), втрачають 7,68 % прибутковості; які володіють середнім капіталом (*X*2=1) – втрачають 5,33 %. Фонди з малим капіталом (*X*3=1) мають невелике (0,72 %) збільшення прибутковості.

У разі орієнтації на швидкий ріст капіталу (*X*4=1) фонд втрачає 15,1 % прибутковості, а повільний ріст (*X*5=1) дозволяє збільшити прибутковість на 2,82 %. У випадку спільного впливу на прибутковість факторів Вид і Мета зміна прибутковості підраховується як сума коефіцієнтів перед фіктивними змінними і їх добутком. Так, для випадку великого капіталу й швидкого росту:

$$\Delta y = -7{,}682 - 15{,}108 - 1{,}979 = -24{,}769\ \%.$$

Для визначення фонду, сприятливого для інвестування, варто врахувати, що в кожному разі сумарна прибутковість визначається за спільного впливу обох факторів. Задачу можна спростити, звівши у формулі (3) облік впливу фіктивних змінних на прибутковість тільки до подвійної суми їх добутків. У результаті роботи інструмента Регресія надбудови Excel Пакет аналізу отримано такий вид функції регресії:

$$Y = -0{,}00034X_А + 6{,}53388X_И - 24{,}7689X_1X_4 - 30{,}4607X_2X_4 - 17{,}4884X_3X_4 -$$
$$- 10{,}5634X_1X_5 + +2{,}1828X_2X_5 + 7{,}3829X_3X_5 \quad (6)$$

Коефіцієнт детермінації $R^2 = 0,678$, що означає, що ця модель точніше пояснює варіації прибутковості. Зміни прибутковості Δy для кожної взаємодії рівнів факторів Вид і Мета збігаються з Δy, обчисленими по залежності (5), як і у випадку великого капіталу й швидкого росту. Отже:

ВИД	МЕТА	ΔY
великий	швидкий	-24,7688
середній		-30,4607
малий		-17,4884
великий	повільний	-10,5633
середній		2,182889
малий		7,382905

Рекомендації щодо вибору фонду, в який варто вкладати кошти, можуть бути такими: фонд із середнім капіталом і націлений на повільний ріст капіталу або фонд із малим капіталом і повільним ростом капіталу. У фонди, орієнтовані на швидкий ріст капіталу, вкладати кошти не рекомендується.

Висновки. Запропоновано модель множинної регресії не тільки на числові, а й на категорійні фактори, у якій категоріальним рівням ставляться у відповідність фіктивні змінні, що набирають значення тільки 1 або 0. Модель дозволяє знайти кількісні оцінки впливу категоріальних рівнів на числову пояснювану змінну. Наведена модель множинної регресії дає можливість прогнозувати значення змінної, що пояснюється, залежно від категоріальних рівнів одного фактора, двох факторів і їх взаємодій.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Васильев А. Н. Научные вычисления в Microsoft Excel : монография / А. Н. Васильев. – Москва : Издательский дом «Вильямс», 2004. – 512 с.
2. Многомерные статистическме методы : учебник / [А. М. Дубров, В. С. Мхитарян, Л. И. Трошин]. – Москва : Финансы и статистика, 2000. – 352 с.
3. Статистика для менеджеров с использованием Microsoft Excel : монография / [Дэвид М. Левин, Дэвид Стефан, Тимоти С. Кребиль, Марк Л. Беренсон]. – Москва : Издательский дом «Вильямс», 2004. – 1312 с.
4. Томашевський В. М. Моделювання систем : монография / В. М. Томашевський. – Київ : Видавнича група BHV, 2005. – 352 с.
5. Цыбрий Л. В. Введение в статистический анализ : учебн. пособ. для вузов / Л. В. Цыбрий. – Днепропетровск : ПГАСА, 2016. – 188 с.

REFERENCES

1. Vasiliev A.N. *Nauchnyye vychisleniya v Microsoft Excel* [Scientific computing in Microsoft Excel]. Moscow : Williams Publishing House, 2004, 512 p. (in Russian).
2. Dubrov A.M., Mkhitaryan V.S and Troshin L.I. *Mnogomernyye statisticheskme metody : uchebnik* [Multicomponent statistical methods : textbook]. *Finansy i statistika* [Finances and statistics], Moscow, 2000, 352 p. (in Russian).
3. Levin D.M., Stefan D., Creebill T.S. and Berenson M.L. *Statistika dlya menedzherov s ispol'zovaniyem Microsoft Excel* [Statistics is for managers from using Microsoft Excel]. Moscow : Williams Publishing House, 2004, 1312 p. (in Russian).
4. Tomashevskyi V.M. *Modelyuvannya system* [Design of the systems]. Kyiv : Publishing Group BHV, 2005, 352 p. (in Russian).
5. Tsybrii L.V. *Vvedeniye v statisticheskiy analiz : uchebn. posobiye dlya vuzov* [Introduction to the statistical analysis: train aid for high schools]. Dnipropetrovsk : PSACEA, 2016, 188 p. (in Russian).

Надійшла до редакції : 13.11.2019.