European Journal of Interdisciplinary Research and Development
Volume-03 June-2022
Website: www.ejird.journalspark.org ISSN (E): 2720-5746

ЗАДАЧА О МНОЖЕСТВЕННОЙ РЕГРЕССИИ

Ёдгаров Самад Жураевич
к.ф-м.н.,доцент ТГПУ им.Низами

Рахматова Севинч Холбековна
магистр

Аннотация
В данной работе с учетом различных аспектов множественной регрессии были отмечены некоторые проблемы, часто возникающие при их практическом использовании

Kalit so'zlar: множественной регрессии, фиктивные переменные, несмещенная оценка, оценка уравнения регрессии, коэффициент детерминации.

Introduction

Рассматривая различные аспекты множественной регрессии, мы отметили некоторые проблемы, часто возникающие при их практическом использовании.

Отметим, что регрессионные модели являются достаточно гибким инструментом, позволяющим, в частности, оценивать влияние качественных признаков (профессия, пол, наличие детей и т.п.) на изучаемую переменную. Это достигается введением в число регрессоров, так называемых фиктивных переменных , принимающих, как правило, значения 1 или 0 в зависимости от наличия или отсутствия соответствующего признака в очередном наблюдении. С формальной точки зрения фиктивные переменные ничем не отличаются от других регрессоров. Однако следует обратить особое внимание на правильное их использование и интерпретацию оценок **(фиктивные переменные).**

Пример. Некоторая фирма занимается продажей молока. В таблице представлены объемы ежемесячных продаж Q (тыс. литров) по различным ценам P (сум за литр). Во время пятого, шестого и седьмого месяцев на одном из предприятий фирмы происходила забастовка.

Месяц	Q	P	Месяц	Q	P
1	98	10,0	8	113	13,0
2	100	11,0	9	116	13,0
3	103	12,5	10	118	13,8
4	105	12,5	11	121	14,2
5	80	14,6	12	123	14,4
6	87	14,6	13	126	15,0
7	94	14,9	14	128	16,1

С помощью регрессий Q на P определите:

а) произошел ли сдвиг свободного члена (константы) во время забастовки по сравнению с обычным режимом;

б) произошел ли сдвиг как константы, так и коэффициента наклона при P.

Решение. Оценим параметры модели множественной регрессии. Определим вектор оценок коэффициентов регрессии. Согласно методу наименьших квадратов, вектор Y(X) получается из выражения:

$\widehat{\boldsymbol{\beta}}_{\mathbf{OLS}} = Y(X) = (X^TX)^{-1}\cdot X^TY$

X – цена (сум за литр);

Y – объем продаж (тыс. литров).

№	X	Y	Z
1	10,0	98	0
2	11,0	100	0
3	12,5	103	0
4	12,5	105	0
5	14,6	80	1
6	14,6	87	1
7	14,9	94	1
8	13,0	113	0
9	13,0	116	0
10	13,8	118	0
11	14,2	121	0
12	14,4	123	0
13	15,0	126	0
14	16,1	128	0

К матрице с переменными X_j добавляем единичный столбец:

$$\text{Матрица X} = \begin{pmatrix} 1 & 10 \\ 1 & 11 \\ 1 & 12,5 \\ 1 & 12,5 \\ 1 & 14,6 \\ 1 & 14,6 \\ 1 & 14,9 \\ 1 & 13 \\ 1 & 13 \\ 1 & 13,8 \\ 1 & 14,2 \\ 1 & 14,4 \\ 1 & 15 \\ 1 & 16.1 \end{pmatrix}; \quad \text{Матрица Y} = \begin{pmatrix} 98 \\ 100 \\ 103 \\ 105 \\ 80 \\ 87 \\ 94 \\ 113 \\ 116 \\ 118 \\ 121 \\ 123 \\ 126 \\ 128 \end{pmatrix}$$

$$\text{Матрица } X^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 10 & 11 & 12,5 & 12,5 & 14,6 & 14,6 & 14,9 & 13 & 13 & 13,8 & 14,2 & 14,4 & 15 & 16,1 \end{pmatrix}$$

$$\text{Умножаем матрицы, } (X^TX) = \begin{pmatrix} 14 & 189,6 \\ 189,6 & 2603,48 \end{pmatrix}$$

$$\text{Находим обратную матрицу } (X^TX)^{-1}: \quad (X^TX)^{-1} = \begin{pmatrix} 5,201 & -0,379 \\ -0,379 & 0,028 \end{pmatrix}$$

$$\text{Умножаем матрицы, } (X^TY) = \begin{pmatrix} 1512 \\ 20564,4 \end{pmatrix}$$

Вектор оценок коэффициентов регрессии равен:

$$Y(X) = \begin{pmatrix} 5,201 & -0,379 \\ -0,379 & 0,028 \end{pmatrix} \cdot \begin{pmatrix} 1512 \\ 20564,4 \end{pmatrix} = \begin{pmatrix} 74,819 \\ 2,45 \end{pmatrix}$$

Уравнение регрессии (оценка уравнения регрессии):

$$\mathbf{Y(X) = 74,8192 + 2,4501 \cdot X} \qquad \mathbf{(1)}$$

Для несмещенной оценки дисперсии проделаем следующие вычисления:

Несмещенная ошибка: ε = Y–Y(x)

№	X	Y	Y(x)	ε = Y - Y(x)	ε^2	$(Y-Ycp)^2$	|ε : Y|
1	10,0	98	99,32	-1,32	1,742	100	0,0135
2	11,0	100	101,77	-1,77	3,132	64	0,0177
3	12,5	103	105,445	-2,445	5,978	25	0,0237
4	12,5	105	105,445	-0,445	0,198	9	0,00424
5	14,6	80	110,59	-30,59	935,752	784	0,382
6	14,6	87	110,59	-23,59	556,491	441	0,271
7	14,9	94	111,325	-17,325	300,158	196	0,184
8	13,0	113	106,67	6,33	40,069	25	0,056
9	13,0	116	106,67	9,33	87,049	64	0,0804
10	13,8	118	108,63	9,37	87,797	100	0,0794
11	14,2	121	109,61	11,39	129,731	169	0,0941
12	14,4	123	110,1	12,9	166,409	225	0,105
13	15,0	126	111,57	14,43	208,223	324	0,115
14	16,1	128	114,265	13,735	188,646	400	0,107
Σ	**189,6**	**1512**			**2711,375**	**2926**	**1,534**
Σ/n	**13,5429**	**108**					

Средняя ошибка аппроксимации:

$$A = \frac{\sum|\varepsilon: Y|}{n} \cdot 100\% = \frac{1,534}{14} \cdot 100\% \approx 10,95\%$$

Оценка дисперсии равна: $s_e^2=(Y-Y(X))^T(Y-Y(X))=2711,375$

Несмещенная оценка дисперсии равна:

$$s^2 = \frac{1}{n-m-1} \cdot s_e^2 = \frac{1}{14-1-1} \cdot 2711,375 = 225,9479$$

Оценка среднеквадратичного отклонения (*стандартная ошибка для оценки Y*):

$$S = \sqrt{s^2} = \sqrt{225,9479} \approx 15,032$$

Найдем оценку ковариационной матрицы вектора $\hat{V}(\hat{\beta}_{OLS}) = S^2 \bullet (X^TX)^{-1}$:

$$\hat{V}(\hat{\beta}_{OLS}) = 225,9479 \cdot \begin{pmatrix} 5,201 & -0,379 \\ -0,379 & 0,028 \end{pmatrix} = \begin{pmatrix} 1175,186 & -85,584 \\ -85,584 & 6,319 \end{pmatrix}$$

Дисперсии параметров модели определяются соотношением $S^2_i = K_{ii}$, т.е. это элементы, лежащие на главной диагонали:

$S_{b0} = \sqrt{1175,186} \approx 34,281$

$S_{b1} = \sqrt{6,319} \approx 2,514$

Тесноту совместного влияния факторов на результат оценивает индекс множественной корреляции. Множественный коэффициент корреляции:

$$R = \sqrt{1 - \frac{s_e^2}{\sum(y_i - \bar{y})^2}} = \sqrt{1 - \frac{2711,375}{2926}} = \sqrt{0,07335} \approx 0,27083$$

Связь между признаком Y и факторами X низкая.

Коэффициент детерминации:

$$R^2 = 0,27083^2 = 0,07335$$

Число *v = n - m - 1* называется числом степеней свободы. Считается, что при оценивании множественной линейной регрессии для обеспечения статистической надежности требуется, чтобы число наблюдений, по крайней мере, в 3 раза превосходило число оцениваемых параметров.

t-статистика:

$$T_{табл}\ (n\text{-}m\text{-}1;\alpha/2) = (12\ ;\ 0,025) = 2,179$$

$$t_i = \frac{b_i}{S_{bi}}$$

$$|t_0| = \left|\frac{74,819}{34,281}\right| \approx 2,183 > 2,179$$

Статистическая значимость коэффициента регрессии b_0 подтверждается.

$$|t_1| = \left|\frac{2,45}{2,514}\right| \approx 0,975 < 2,179$$

Статистическая значимость коэффициента регрессии b_1 не подтверждается.

F-статистика. Критерий Фишера.

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m} = \frac{0,07335}{1 - 0,07335} \cdot \frac{14 - 1 - 1}{1} \approx 0,95$$

Табличное значение при степенях свободы $k_1=1$ и $k_2=n–m–1=14–1–1=12$,

$$F_{табл}\ (1;12) = 4,75$$

Поскольку фактическое значение $F_{набл} < F_{табл}$, то коэффициент детерминации статистически не значим и уравнение регрессии статистически ненадежно.

При увеличения цены молоко на 1 сум за литр приводит к увеличению объема продажи в среднем на 2,45 тыс. литров. Статистическая значимость уравнения проверена с помощью коэффициента детерминации и критерия Фишера. Установлено, что в исследуемой ситуации 7,34% общей вариабельности Y объясняется изменением факторов X_j. Установлено также, что параметры модели статистически не значимы.

Полученная уравнения не учитывает влияние признака — фактора «забастовка».

Для ее учета введем в регрессионную модель фиктивную (бинарную) переменную $\mathbf{Z_1}$,

где, $z_{i1} = \begin{cases} 1, & \text{если забастовка совершалась,} \\ 0, & \text{если обычный рабочий режим.} \end{cases}$

К матрице с переменными X добавляем столбец Z и получим матрицу X*

$$\text{Матрица } X^{*T} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 10 & 11 & 12,5 & 12,5 & 14,6 & 14,6 & 14,9 & 13 & 13 & 13,8 & 14,2 & 14,4 & 15 & 16,1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\text{Умножаем матрицы, } (X^{*T} \cdot X^*) = \begin{pmatrix} 14 & 189,6 & 3 \\ 189,6 & 2603,48 & 44,1 \\ 3 & 44,1 & 3 \end{pmatrix}$$

Находим обратную матрицу $(X^{*T} \cdot X^*)^{-1}$:

$$(X^{*T}\cdot X^{*})^{-1} = \begin{pmatrix} 5{,}801 & -0{,}432 & 0{,}545 \\ -0{,}432 & 0{,}028 & -0{,}0481 \\ 0{,}545 & -0{,}0481 & 0{,}495 \end{pmatrix}$$

$$\text{Умножаем матрицы, } (X^{*T}\cdot Y) = \begin{pmatrix} 1512 \\ 20564{,}4 \\ 261 \end{pmatrix}$$

Вектор оценок коэффициентов регрессии равен:

$$Y(X,Z) = \begin{pmatrix} 5{,}801 & -0{,}432 & 0{,}545 \\ -0{,}432 & 0{,}028 & -0{,}0481 \\ 0{,}545 & -0{,}0481 & 0{,}495 \end{pmatrix} \cdot \begin{pmatrix} 1512 \\ 20564{,}4 \\ 261 \end{pmatrix} = \begin{pmatrix} 35{,}861 \\ 5{,}887 \\ -35{,}397 \end{pmatrix}$$

Уравнение регрессии (оценка уравнения регрессии):

$$\mathbf{Y(X,Z)=35{,}8612+5{,}8868\cdot X-35{,}3969\cdot Z} \qquad \mathbf{(2)}$$

Для несмещенной оценки дисперсии проделаем следующие вычисления:

Несмещенная ошибка: ε = Y–Y(x)

№	X	Z	Y	Y(X,Z)	ε = Y - Y(x,z)	ε^2	$(Y-Ycp)^2$	|ε : Y|
1	10,0	0	98	94,729	3,271	10,699	100	0,033
2	11,0	0	100	100,616	-0,616	0,379	64	0,006
3	12,5	0	103	109,446	-6,446	41,551	25	0,063
4	12,5	0	105	109,446	-4,446	19,767	9	0,042
5	14,6	1	80	86,4113	-6,411	41,105	784	0,080
6	14,6	1	87	86,4113	0,589	0,347	441	0,007
7	14,9	1	94	88,1774	5,823	33,903	196	0,062
8	13,0	0	113	112,389	0,611	0,373	25	0,005
9	13,0	0	116	112,389	3,611	13,037	64	0,031
10	13,8	0	118	117,099	0,901	0,812	100	0,008
11	14,2	0	121	119,454	1,546	2,392	169	0,013
12	14,4	0	123	120,631	2,369	5,613	225	0,019
13	15,0	0	126	124,163	1,837	3,375	324	0,015
14	16,1	0	128	130,638	-2,638	6,961	400	0,021
Σ	**189,6**	**3**	**1512**	**1512**		**180,313**	**2926**	**0,405**
Σ/n	**13,5429**		**108**	**108**				

Средняя ошибка аппроксимации:

$$A = \frac{\sum|\varepsilon: Y|}{n}\cdot 100\% = \frac{0{,}405}{14}\cdot 100\% \approx 2{,}89\%$$

Оценка дисперсии равна: $s_e^2=(Y-Y(X))^T(Y-Y(X))=180{,}313$

Несмещенная оценка дисперсии равна:

$$s^2 = \frac{1}{n-m-1}\cdot s_e^2 = \frac{1}{14-2-1}\cdot 180{,}313 = 16{,}3921$$

Оценка среднеквадратичного отклонения (*стандартная ошибка для оценки Y*): $S = \sqrt{s^2} = \sqrt{16{,}3921} \approx 4{,}049$

Найдем оценку ковариационной матрицы вектора $\hat{V}(\hat{\beta}_{OLS}) = S^2 \bullet (X^{*T}\cdot X^{*})^{-1}$:

$$\hat{V}(\hat{\beta}_{OLS}) = 16{,}3921 \cdot \begin{pmatrix} 5{,}801 & -0{,}432 & 0{,}545 \\ -0{,}432 & 0{,}028 & -0{,}0481 \\ 0{,}545 & -0{,}0481 & 0{,}495 \end{pmatrix} = \begin{pmatrix} 95{,}087 & -7{,}076 & 8{,}931 \\ -7{,}076 & 0{,}535 & -0{,}788 \\ 8{,}931 & -0{,}788 & 8{,}114 \end{pmatrix}$$

Дисперсии параметров модели определяются соотношением $S^2_i = K_{ii}$, т.е. это элементы, лежащие на главной диагонали:

$S_{b0} = \sqrt{95{,}087} \approx 9{,}751$

$S_{b1} = \sqrt{0{,}535} \approx 0{,}731$

$S_{b2} = \sqrt{8{,}114} \approx 2{,}849$

Тесноту совместного влияния факторов на результат оценивает индекс множественной корреляции. Множественный коэффициент корреляции:

$$R = \sqrt{1 - \frac{s_e^2}{\sum(y_i - \bar{y})^2}} = \sqrt{1 - \frac{180{,}313}{2926}} = \sqrt{0{,}938} \approx 0{,}9687$$

Связь между признаком Y и факторами X сильная.

Коэффициент детерминации: $R^2 = 0{,}9687^2 = 0{,}9384$

Более объективной оценкой является *скорректированный коэффициент детерминации:*

$$\bar{R}^2 = 1 - (1 - 0{,}9384) \cdot \frac{14 - 1}{14 - 2 - 1} \approx 0{,}927$$

Число *v = n - m - 1* называется числом степеней свободы. Считается, что при оценивании множественной линейной регрессии для обеспечения статистической надежности требуется, чтобы число наблюдений, по крайней мере, в 3 раза превосходило число оцениваемых параметров.

t-статистика:

$T_{табл}$ (n-m-1;α/2) = (11 ; 0,025) = 2,201

$$t_i = \frac{b_i}{S_{bi}}$$

$$|t_0| = \left|\frac{35{,}861}{9{,}751}\right| \approx 3{,}678 > 2{,}201$$

Статистическая значимость коэффициента регрессии b_0 подтверждается.

$$|t_1| = \left|\frac{5{,}887}{0{,}731}\right| \approx 8{,}049 > 2{,}201$$

Статистическая значимость коэффициента регрессии b_1 подтверждается.

$$|t_2| = \left|\frac{-35{,}397}{2{,}849}\right| \approx 12{,}426 > 2{,}201$$

Статистическая значимость коэффициента регрессии b_2 подтверждается.

F-статистика. Критерий Фишера.

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m} = \frac{0{,}9384}{1 - 0{,}9384} \cdot \frac{14 - 2 - 1}{2} \approx 83{,}751$$

Табличное значение при степенях свободы $k_1=2$ и $k_2=n–m–1=14–2–1=11$,

$F_{табл}$ (2;11) = 3,98

Поскольку фактическое значение $F_{набл} > F_{табл}$, то коэффициент детерминации статистически значим и уравнение регрессии статистически надежно (т.е. коэффициенты b_i совместно значимы).

Следовательно, по имеющимся данным влияние фактора «забастовка» оказалось существенным, и у нас есть основания считать, что регрессионная модель объема реализации молока от цены существенно различается по сравнению при обычном рабочем режиме и когда совершалась забастовка.

Литература

1. Крамер Г. Математические методы статистики. М.: Мир, 1975. – 648 с.
2. Берндт, Э. Р. Практика эконометрики: классика и современность : учебник / Э. Р. Берндт. – М. : ЮНИТИ-ДАНА, 2005. – 863 с.