Применение множественной регрессии при решении экономических задач

БЕЛОГЛАЗОВА ЮЛИЯ, ДС-01 МЭ

По 20 предприятиям региона изучается зависимость выработки продукции на одного работника *y* (тыс. руб.) от ввода в действие новых основных фондов x_1 (% от стоимости фондов на конец года) и от удельного веса рабочих высокой квалификации в общей численности рабочих x_2 (%)

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	7,0	3,9	10,0	11	9,0	6,0	21,0
2	7,0	3,9	14,0	12	11,0	6,4	22,0
3	7,0	3,7	15,0	13	9,0	6,8	22,0
4	7,0	4,0	16,0	14	11,0	7,2	25,0
5	7,0	3,8	17,0	15	12,0	8,0	28,0
6	7,0	4,8	19,0	16	12,0	8,2	29,0
7	8,0	5,4	19,0	17	12,0	8,1	30,0
8	8,0	4,4	20,0	18	12,0	8,5	31,0
9	8,0	5,3	20,0	19	14,0	9,6	32,0
10	10,0	6,8	20,0	20	14,0	9,0	36,0

Требуется:

1. Построить линейную модель множественной регрессии. Ранжировать факторы по степени их влияния на результат
2. Найти коэффициенты парной, частной и множестенной корреляции. Проанализировать их.
3. Найти скорректированный коэффициент множественной детерминации. Сравнить его с нескорректированным (общим) коэффициентом детерминации.
4. С помощью *F* -критерия Фишера оценить статистическую надежность уравнения регрессии и коэффициента детерминации
5. С помощью *t* -критерия оценить статистическую значимость коэффициентов чистой регрессии.
6. С помощью частных *F* -критериев Фишера оценить целесообразность включения в уравнение множественной регрессии фактора *x1* после *x2* и фактора *x2* после *x1* .
7. Составить уравнение линейной парной регрессии, оставив лишь один значащий фактор.

Найдем средние квадратические отклонения

1

$$\sigma_y = \sqrt{\overline{y^2} - \overline{y}^2} = \sqrt{97,9 - 9,6^2} = \sqrt{5,74} = 2,396;$$

$$\sigma_{x_1} = \sqrt{\overline{x_1^2} - \overline{x}_1^2} = \sqrt{41,887 - 6,19^2} = \sqrt{3,571} = 1,890;$$

$$\sigma_{x_2} = \sqrt{\overline{x_2^2} - \overline{x}_2^2} = \sqrt{541,4 - 22,3^2} = \sqrt{44,11} = 6,642.$$

Найдем параметры линейного уравнения множественной регрессии

Коэффициенты корреляции → Параметры

$$r_{yx_1} = \frac{\operatorname{cov}(y, x_1)}{\sigma_y \cdot \sigma_{x_1}} = \frac{63,815 - 6,19 \cdot 9,6}{1,890 \cdot 2,396} = 0,970;$$

$$r_{yx_2} = \frac{\operatorname{cov}(y, x_2)}{\sigma_y \cdot \sigma_{x_2}} = \frac{229,05 - 22,3 \cdot 9,6}{6,642 \cdot 2,396} = 0,941;$$

$$r_{x_1x_2} = \frac{\operatorname{cov}(x_1, x_2)}{\sigma_{x_1} \cdot \sigma_{x_2}} = \frac{149,87 - 6,19 \cdot 22,3}{1,890 \cdot 6,642} = 0,943$$

$$b_1 = \frac{\sigma_y}{\sigma_{x_1}} \cdot \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{1 - r_{x_1x_2}^2} = \frac{2,396}{1,890} \cdot \frac{0,970 - 0,941 \cdot 0,943}{1 - 0,943^2} = 0,946;$$

$$b_2 = \frac{\sigma_y}{\sigma_{x_2}} \cdot \frac{r_{yx_2} - r_{yx_1} r_{x_1x_2}}{1 - r_{x_1x_2}^2} = \frac{2,396}{6,642} \cdot \frac{0,941 - 0,970 \cdot 0,943}{1 - 0,943^2} = 0,0856;$$

$$a = \overline{y} - b_1\overline{x}_1 - b_2\overline{x}_2 = 9,6 - 0,946 \cdot 6,19 - 0,0856 \cdot 22,3 = 1,835.$$

Уравнение множественной регрессии

$$\hat{y} = 1{,}835 + 0{,}946 \cdot x_1 + 0{,}0856 \cdot x_2.$$

При увеличении ввода в действие основных фондов на 1% (при неизменном уровне удельного веса рабочих высокой квалификации) выработка продукции на одного рабочего увеличивается в среднем на 0,946 тыс. руб

При увеличении удельного веса рабочих высокой квалификации в общей численности рабочих на 1% (при неизменном уровне ввода в действие новых основных фондов) выработка продукции на одного рабочего увеличивается в среднем на 0,086 тыс. руб.

№	y	x_1	x_2	$\hat{y}$	$y-\hat{y}$	$(y-\hat{y})^2$	A_i, %
1	7,0	3,9	10,0	6,380	0,620	0,384	8,851
2	7,0	3,9	14,0	6,723	0,277	0,077	3,960
3	7,0	3,7	15,0	6,619	0,381	0,145	5,440
4	7,0	4,0	16,0	6,989	0,011	0,000	0,163
5	7,0	3,8	17,0	6,885	0,115	0,013	1,643
6	7,0	4,8	19,0	8,002	-1,002	1,004	14,317
7	8,0	5,4	19,0	8,570	-0,570	0,325	7,123
8	8,0	4,4	20,0	7,709	0,291	0,084	3,633
9	8,0	5,3	20,0	8,561	-0,561	0,315	7,010
10	10,0	6,8	20,0	9,980	0,020	0,000	0,202
11	9,0	6,0	21,0	9,309	-0,309	0,095	3,429
12	11,0	6,4	22,0	9,773	1,227	1,507	11,158
13	9,0	6,8	22,0	10,151	-1,151	1,325	12,789
14	11,0	7,2	25,0	10,786	0,214	0,046	1,944
15	12,0	8,0	28,0	11,800	0,200	0,040	1,668
16	12,0	8,2	29,0	12,075	-0,075	0,006	0,622
17	12,0	8,1	30,0	12,066	-0,066	0,004	0,547
18	12,0	8,5	31,0	12,530	-0,530	0,280	4,413
19	14,0	9,6	32,0	13,656	0,344	0,118	2,459
20	14,0	9,0	36,0	13,431	0,569	0,324	4,067
Сумма	192	123,8	446	191,992	0,008	6,093	95,437
Ср. знач.	9,6	6,19	22,3	9,6	–	0,305	4,77

Результаты

Средняя ошибка аппроксимации:

$$\overline{A} = \frac{1}{n}\sum\left|\frac{y-\hat{y}}{y}\right| \cdot 100\% = \frac{95,437\%}{20} = 4,77\%.$$

Качество модели, исходя из относительных отклонений по каждому наблюдению, признается хорошим, т.к. средняя ошибка аппроксимации не превышает 10%

Введение в действие основных фондов влияет на результат больше, чем увеличение числа рабочих высокой квалификации

Коэффициенты корреляции

2

$$r_{yx_1} = 0{,}970;\ r_{yx_2} = 0{,}941;\ r_{x_1x_2} = 0{,}943.$$

Они указывают на весьма сильную связь каждого фактора с результатом, а также высокую межфакторную зависимость (факторы *x1* и *x2* явно коллинеарны)

При такой сильной межфакторной зависимости рекомендуется один из факторов исключить из рассмотрения.

Частные коэффициенты корреляции

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_2}^2\right) \cdot \left(1 - r_{x_1x_2}^2\right)}} = \frac{0{,}970 - 0{,}941 \cdot 0{,}943}{\sqrt{\left(1 - 0{,}941^2\right) \cdot \left(1 - 0{,}943^2\right)}} = 0{,}734;$$

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_1}^2\right) \cdot \left(1 - r_{x_1x_2}^2\right)}} = \frac{0{,}941 - 0{,}970 \cdot 0{,}943}{\sqrt{\left(1 - 0{,}970^2\right) \cdot \left(1 - 0{,}943^2\right)}} = 0{,}325.$$

Если сравнить коэффициенты парной и частной корреляции, то можно увидеть, что из-за высокой межфакторной зависимости коэффициенты парной корреляции дают завышенные оценки тесноты связи.

Коэффициент множественной корреляции

Остаточная дисперсия:

$$\sigma^2_{ост} = \frac{\sum (y - \hat{y})^2}{n} = \frac{6,093}{20} = 0,305.$$

$$R_{yx_1x_2} = \sqrt{1 - \frac{\sigma^2_{ост}}{\sigma^2_y}} = \sqrt{1 - \frac{0,305}{5,74}} = 0,973;$$

Коэффициент множественной корреляции указывает на весьма сильную связь всего набора факторов с результатом.

Коэффициент множественной детерминации

3

Скорректированный коэффициент множественной детерминации

$$R^2_{yx_1x_2} = 0{,}947$$

Нескорректированный коэффициент множественной детерминации

$$\widetilde{R}^2 = 1-(1-R^2)\frac{(n-1)}{(n-m-1)} = 1-(1-0{,}947)\frac{20-1}{20-2-1} = 0{,}941$$

Оба коэффициента указывают на весьма высокую (более 94%) детерминированность результата *y* в модели факторами *x1* и 2 *x* .

F -критерий Фишера

4

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m}.$$

Подтверждается статистическая значимость всего уравнения и показателя тесноты связи – к-та детерминации

$$F_{\text{факт}} = \frac{0{,}973^2}{1-0{,}973^2} \cdot \frac{20-2-1}{2} = 151{,}88.$$

$$F_{\text{факт}} = 151{,}88 > F_{\text{табл}} = 3{,}59 \quad (\text{при} \quad n = 20)$$

5

t -критерий

Стандартные ошибки коэффициентов регрессии

$$m_{b_1} = \frac{\sigma_y \cdot \sqrt{1-R^2_{yx_1x_2}}}{\sigma_{x_1} \cdot \sqrt{1-r^2_{x_1x_2}}} \cdot \frac{1}{\sqrt{n-3}} = \frac{2,396 \cdot \sqrt{1-0,973^2}}{1,890 \cdot \sqrt{1-0,943^2}} \cdot \frac{1}{\sqrt{20-3}} = 0,2132;$$

$$m_{b_2} = \frac{\sigma_y \cdot \sqrt{1-R^2_{yx_1x_2}}}{\sigma_{x_2} \cdot \sqrt{1-r^2_{x_1x_2}}} \cdot \frac{1}{\sqrt{n-3}} = \frac{2,396 \cdot \sqrt{1-0,973^2}}{6,642 \cdot \sqrt{1-0,943^2}} \cdot \frac{1}{\sqrt{20-3}} = 0,0607.$$

Фактические значения *t* -критерия Стьюдента:

$$t_{b_1} = \frac{b_1}{m_{b_1}} = \frac{0,946}{0,2132} = 4,44, \; t_{b_2} = \frac{b_2}{m_{b_2}} = \frac{0,0856}{0,0607} = 1,41.$$

Табличное значение критерия при уровне значимости a = 0,05 и числе степеней свободы *k* =17 составит 2,11.

Таким образом, признается статистическая значимость параметра *b1* и случайная природа формирования параметра b2

Частные *F* -критерии Фишера

6

$$F_{x_1}=\frac{R^2_{yx_1x_2}-R^2_{yx_2}}{1-R^2_{yx_1x_2}}\cdot\frac{n-m-1}{1};\quad F_{x_2}=\frac{R^2_{yx_1x_2}-R^2_{yx_1}}{1-R^2_{yx_2x_2}}\cdot\frac{n-m-1}{1}.$$

$$R^2_{yx_1}=r^2_{yx_1}=0{,}970^2=0{,}941;$$

$$R^2_{yx_2}=r^2_{yx_2}=0{,}941^2=0{,}885.$$

$$F_{x_1}=\frac{0{,}947-0{,}885}{1-0{,}947}\cdot\frac{20-2-1}{1}=19{,}89;$$

$$F_{x_2}=\frac{0{,}947-0{,}941}{1-0{,}947}\cdot\frac{20-2-1}{1}=1{,}924.$$

$$F_{\text{табл}}\left(\alpha=0{,}05;\ k_1=1;\ k_2=17\right)=4{,}45$$

Следовательно, включение в модель фактора *x2* после того, как в модель включен фактор *x1* статистически нецелесообразно: прирост факторной дисперсии за счет дополнительного признака *x2* оказывается незначительным

Уравнение линейной парной регрессии

7

$$\hat{y}_{x_1} = \alpha + \beta x_1 .$$

$$\beta = \frac{\operatorname{cov}(y,\ x_1)}{\sigma_{x_2}} = \frac{63,815 - 6,19 \cdot 9,6}{3,571} = 1,23;$$

$$\alpha = \bar{y} - \beta \cdot \bar{x} = 9,6 - 1,23 \cdot 6,19 = 1,99 .$$

$$\hat{y}_{x_1} = 1,99 + 1,23 \cdot x_1, \qquad r^2_{yx_1} = 0,941 .$$

Общий вывод состоит в том, что множественная модель с факторами *x1* и *x2* содержит неинформативный фактор *x2* .
Если исключить фактор *x2* , то можно ограничиться уравнением парной регрессии