Обработка экспериментальных данных

Лекция 7: Многофакторная регрессия

Преподаватель: Аникеева Александра Евгеньевна

Отбор факторных признаков для включения в модель

Находят коэффициенты парной корреляции $r_{X_1X_2}$, $r_{X_1X_3}$, $r_{X_2X_3}$.

Если их значения меньше $0{,}8$, их включают в модель.

Пусть , $r_{X_2X_3} = 0{,}9$. Следовательно один из признаков - X_2 или X_3 – нужно исключить из модели

Рассчитать коэффициенты парной корреляции между каждым из факторов (в нашем примере X_2 и X_3)и результативным признаком Y: r_{YX_2} и r_{YX_3} .

Если, получилось, что $r_{YX_2} > r_{YX_3}$, то исключают признак X_3

Если $|r_{X_iX_j}| > 1 - \frac{3(1-r^2_{X_iX_j})}{\sqrt{n-1}}$, то один из факторов можно исключить

С помощью t-критерия Стьюдента:

$$t_{набл} = \sqrt{\frac{r^2_{YX_j}}{1-r^2_{YX_j}}(n-2)}$$

Если $t_{набл} > t_{крит}$, $(\alpha\ ;\ n\text{-}2)$ то с вероятностью $1\text{-}\alpha$ можно утверждать о значимости межфакторного коэффициента корреляции и, следовательно, о значимости факторного признака X_j. Он включается в модель.

Двухфакторная регрессионная модель

$$y^m = a_1 x_1 + a_2 x_2 + b$$

$$a_1 \sum_{i=1}^{n} x_{1i} + a_2 \sum_{i=1}^{n} x_{2i} + nb = \sum_{i=1}^{n} y_i$$

$$a_1 \sum_{i=1}^{n} x_{1i}^2 + a_2 \sum_{i=1}^{n} x_{1i} x_{2i} + b \sum_{i=1}^{n} x_{1i} = \sum_{i=1}^{n} y_i x_{1i}$$

$$a_1 \sum_{i=1}^{n} x_{1i} x_{2i} + a_2 \sum_{i=1}^{n} x_{2i}^2 + b \sum_{i=1}^{n} x_{2i} = \sum_{i=1}^{n} y_i x_{2i}$$

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i; \quad \overline{y^2} = \frac{1}{n} \sum_{i=1}^{n} y_i^2;$$

$$S_y^2 = \overline{y^2} - \left(\overline{y}\right)^2$$

$$S_{x_i}^2 = \overline{x_i^2} - \left(\overline{x_i}\right)^2, \; i = 1,2$$

$$a_1 = \frac{r_{YX_1} - r_{X_1X_2} r_{YX_2}}{1 - r_{X_1X_2}^2} \cdot \frac{S_y}{S_{X_1}}, \qquad a_2 = \frac{r_{YX_2} - r_{X_1X_2} r_{YX_1}}{1 - r_{X_1X_2}^2} \cdot \frac{S_y}{S_{X_2}}$$

$$\overline{x_j} = \frac{1}{n} \sum_{i=1}^{n} x_{ji}; \; j = 1,2$$

$$b = \overline{y} - a_1 \overline{x_1} - a_2 \overline{x_2}$$

Множественной линейной корреляционной связи

$$R_{Y/X_1X_2} = \sqrt{\frac{r_{YX_1}^2 + r_{YX_2}^2 - 2r_{YX_1} \cdot r_{YX_2} \cdot r_{X_1X_2}}{1 - r_{X_1X_2}^2}}$$

Свойства множественного коэффициента корреляции:

1) $0 \le R \le 1$
2) Если $R=0$, то линейная корреляционная связь между X_1 и X_2 и Y отсутствует, хотя между ними может существовать нелинейная или функциональная зависимость.
3) Если $R=1$, то, между X_1 и X_2 и Y существует линейная функциональная зависимость.-
4) При небольшом объеме выборки величину множественного коэффициента корректируют

$$\hat{R}_{Y/X_1X_2} = \sqrt{1 - (1 - R_{Y/X_1X_2}^2)\frac{n-1}{n-k-1}}$$

Здесь $k=2$ – число факторных признаков.

Для коэффициента множественной корреляции определяют среднеквадратическую ошибку

$$S_R = \frac{1}{\sqrt{n-1}}$$

Если $\frac{R}{S_R} > 3$, то с вероятностью 0,99 можно считать R значимым.

Проверка адекватности модели множественной линейной корреляции

Вычисляют статистику

$$F_{набл} = \frac{(n-k-1)\cdot(R_{Y/X_1X_3})^2}{k\left(1-(R_{Y/X_1X_3})^2\right)}$$

Здесь n– объем выборки, k – число факторных признаков

При заданном уровне значимости α и степенях свободы $\nu_1=k$ и $\nu_2=n-k-1$ по таблице критических точек распределения Фишера находят критическое значение $F_{крит}$. Если $F_{набл} > F_{крит}$, то уравнение регрессии согласуется с опытными данными.

Адекватность модели множественной корреляции можно определить по средней ошибке аппроксимации

$$\bar{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - y_i^m}{y_i}\right|100\%$$

Коэффициент эластичности: $K_э = f'(x)\frac{\bar{x}}{\bar{y}}$

Пример

Имеются следующие показатели по предприятиям отрасли:

№ предприятия	Стоимость промышленно-производственных основных фондов, тыс. руб	Валовая продукция в оптовых ценах предприятия, тыс. руб.	Среднесписочная численность промышленно – производственного персонала, чел.	Среднесписочная численность рабочих, чел.
1	4999	5349	420	331
2	6929	6882	553	486
3	6902	7046	570	498
4	10097	7248	883	789
5	8097	5256	433	359
6	11116	14090	839	724
7	4880	3525	933	821
8	7355	5431	526	428
9	10066	7680	676	607
10	7884	8226	684	619

$$\overline{y} = \frac{1}{10}\sum_{i=1}^{10} y_i = \frac{78325}{10} = 7832{,}5; \quad \overline{y^2} = \frac{1}{10}\sum_{i=1}^{10} y_i^2 = \frac{653107157}{10} = 65310715{,}7$$

$$S_y^2 = \overline{y^2} - \left(\overline{y}\right)^2 = 65310715{,}7 - 7832{,}5^2 = 3962659{,}45 \qquad S_y = \sqrt{3962659{,}45} = 1990{,}642974$$

$$\overline{x_1} = \frac{1}{10}\sum_{i=1}^{10} x_{1i} = \frac{70733}{10} = 7073{,}3; \qquad S_{x_1}^2 = \overline{x_1^2} - \left(\overline{x_1}\right)^2 = 57287784{,}3 - 7073{,}3^2 = 7256211{,}41$$

$$\overline{x_1^2} = \frac{1}{10}\sum_{i=1}^{10} x_{1i}^2 = \frac{572877843}{10} = 57287784{,}3 \qquad S_{x_1} = \sqrt{7256211{,}41} = 2693{,}735587$$

$$\overline{x_2} = \frac{1}{10}\sum_{i=1}^{10} x_{2i} = \frac{6517}{10} = 651{,}7; \qquad S_{x_2}^2 = \overline{x_2^2} - \left(\overline{x_2}\right)^2 = 455020{,}5 - 651{,}7^2 = 30307{,}61$$

$$\overline{x_2^2} = \frac{1}{10}\sum_{i=1}^{10} x_{2i}^2 = \frac{4550205}{10} = 455020{,}5 \qquad S_{x_2} = \sqrt{30307{,}61} = 174{,}09081$$

$$\overline{x_3} = \frac{1}{10}\sum_{i=1}^{10} x_{3i} = \frac{5662}{10} = 566{,}2; \qquad S_{x_3}^2 = \overline{x_3^2} - \left(\overline{x_3}\right)^2 = 347817{,}4 - 566{,}2^2 = 27234{,}96$$

$$\overline{x_3^2} = \frac{1}{10}\sum_{i=1}^{10} x_{3i}^2 = \frac{3478174}{10} = 347817{,}4 \qquad S_{x_3} = \sqrt{27234{,}96} = 165{,}030179$$

$$\overline{yx_1} = \frac{1}{10}\sum_{i=1}^{10} y_i x_{1i} = \frac{594729518}{10} = 59472951{,}8$$

$$\overline{yx_2} = \frac{1}{10}\sum_{i=1}^{10} y_i x_{2i} = \frac{52232475}{10} = 5223247{,}5$$

$$\overline{yx_3} = \frac{1}{10}\sum_{i=1}^{10} y_i x_{3i} = \frac{45525377}{10} = 4552537{,}7$$

$$\overline{x_1 x_2} = \frac{1}{10}\sum_{i=1}^{10} x_{1i} x_{2i} = \frac{47529683}{10} = 4752968{,}3$$

$$\overline{x_1 x_3} = \frac{1}{10}\sum_{i=1}^{10} x_{1i} x_{3i} = \frac{41402962}{10} = 4140296{,}2$$

$$\overline{x_2 x_3} = \frac{1}{10}\sum_{i=1}^{10} x_{2i} x_{3i} = \frac{3976057}{10} = 397605{,}7$$

$$r_{YX_1} = \frac{\overline{yx_1} - \overline{y}\,\overline{x_1}}{S_y S_{x_1}} = \frac{59472951{,}8 - 7832{,}5 \cdot 7073{,}3}{1990{,}642974 \cdot 2693{,}735587} = 0{,}759255 \qquad r^2_{YX_1} = (0{,}759255)^2 = 0{,}576469$$

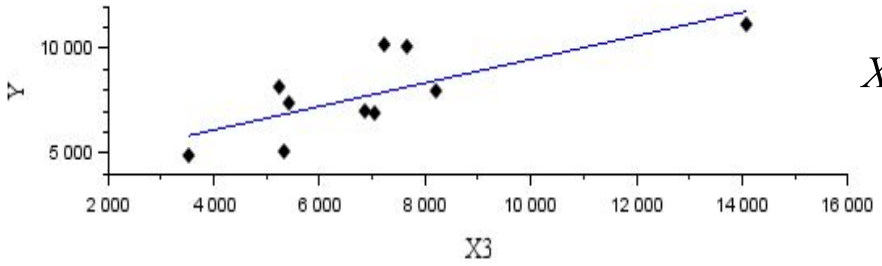
$$r_{YX_2} = \frac{\overline{yx_2} - \overline{y}\,\overline{x_2}}{S_y S_{x_2}} = \frac{5223247{,}5 - 7832{,}5 \cdot 651{,}7}{1990{,}642974 \cdot 174{,}09081} = 0{,}342826 \qquad r^2_{YX2} = (0{,}342826)^2 = 0{,}11753$$

$$r_{YX_3} = \frac{\overline{yx_3} - \overline{y}\,\overline{x_3}}{S_y S_{x_3}} = \frac{4552537{,}7 - 7832{,}5 \cdot 566{,}2}{1990{,}642974 \cdot 165{,}030179} = 0{,}35851 \qquad r^2_{YX_3} = (0{,}35851)^2 = 0{,}128529$$

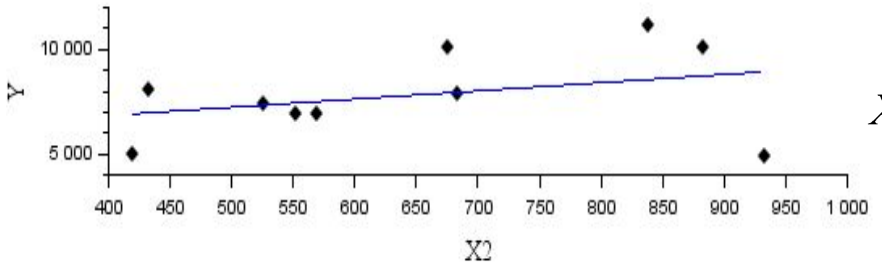
$$r_{X_1X_2} = \frac{\overline{x_1x_2} - \overline{x_1}\,\overline{x_2}}{S_{x_1} S_{x_2}} = \frac{4752968{,}3 - 7073{,}3 \cdot 651{,}7}{2693{,}735587 \cdot 174{,}09081} = 0{,}30557 \qquad r^2_{X_1X_2} = (0{,}30557)^2 = 0{,}093373$$

$$r_{X_1X_3} = \frac{\overline{x_1x_3} - \overline{x_1} \cdot \overline{x_3}}{S_{x_1} S_{x_3}} = \frac{4140296{,}2 - 7073{,}3 \cdot 566{,}2}{2693{,}735587 \cdot 165{,}030179} = 0{,}304565 \qquad r^2_{X_1X_3} = (0{,}304565)^2 = 0{,}09276$$

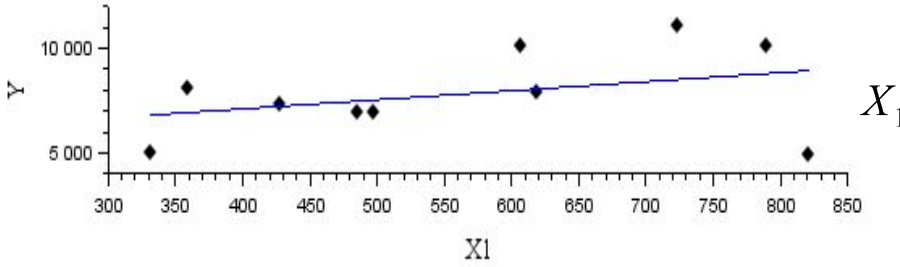
$$r_{X_2X_3} = \frac{\overline{x_2x_3} - \overline{x_2}\,\overline{x_3}}{S_{x_2} S_{x_3}} = \frac{397605{,}7 - 651{,}7 \cdot 566{,}2}{174{,}09081 \cdot 165{,}030179} = 0{,}995925 \qquad r^2_{X_2X_3} = (0{,}995925)^2 = 0{,}991866$$



$$X_3: t_{набл} = \sqrt{\frac{r^2_{YX_3}}{1-r^2_{YX_3}}(n-2)} = \sqrt{\frac{0{,}128529 \cdot 8}{1-0{,}128529}} = 1{,}0862239$$



$$X_2: t_{набл} = \sqrt{\frac{r^2_{YX_2}}{1-r^2_{YX_2}}(n-2)} = \sqrt{\frac{0{,}11753 \cdot 8}{1-0{,}11753}} = 1{,}0322114$$



$$X_1: t_{набл} = \sqrt{\frac{r^2_{YX_1}}{1-r^2_{YX_1}}(n-2)} = \sqrt{\frac{0{,}576469 \cdot 8}{1-0{,}576469}} = 3{,}2998198$$

Поскольку для X_1 $t_{набл} > t_{крит} = 2{,}31$, X_1 включается в модель.

Связь между признаками X_2 и X_3 тесная: $r_{X_2X_3} = 0{,}9959 > 0{,}8$

Поскольку $r_{YX_3} = 0{,}128529 > r_{YX_2} = 0{,}11753$, то признак X_2 исключается из рассмотрения, а признак X_3 - остается.

Множественный коэффициент корреляции

$$R_{Y/X_1X_3} = \sqrt{\frac{r^2_{YX_1} + r^2_{YX_3} - 2r_{YX_1} \cdot r_{YX_3} \cdot r_{X_1X_3}}{1 - r^2_{X_1X_3}}} =$$

$$= \sqrt{\frac{0{,}576469 + 0{,}128529 - 2 \cdot 0{,}759255 \cdot 0{,}35851 \cdot 0{,}304565}{1 - 0{,}09276}} = 0{,}7709$$

$$\hat{R}_{Y/X_1X_2} = \sqrt{1 - (1 - R^2{}_{Y/X_1X_2}) \frac{n-1}{n-k-1}} = \sqrt{1 - (1 - 0{,}7709^2) \cdot \frac{9}{7}} = 0{,}69167$$

Проверим значимость коэффициента корреляции по критерию Стьюдента

$$t_{выб} = \frac{\hat{R}}{S_R} = 3 \cdot 0{,}69167 = 2{,}07502$$

По таблице критических точек распределения Стьюдента при уровне значимости $\alpha=0{,}05$ и числе степеней свободы $k=n-3=7$ находим $t_{крит} = 1{,}89458$.

Так как $t_{выб} > t_{крит}$, делаем вывод, что значим.

Нахождение коэффициентов выбранной зависимости

$$70733\,a_1 + 5662\,a_2 + 10\,b = 78325$$

$$572877843\,a_1 + 41402962\,a_2 + 70733\,b = 594729518$$

$$41402962\,a_1 + 3478174\,a_2 + 5662\,b = 45525377$$

$$y = 0{,}5295093x_1 + 1{,}6922088x_3 + 3129{,}061528$$

Общий индекс детерминации равен

$$R = R^2_{Y/X_1X_3} = 0{,}7709^2 = 0{,}59432172$$

Следовательно, факторные признаки, отобранные в модель, влияют на результативный в пределах 59,43%. Это не очень сильное влияние. Согласно закону Парето степень влияния должна быть не меньше 80%.

Проверка адекватности модели

$$F_{набл} = \frac{(n-k-1)\cdot(R_{Y/X_1X_3})^2}{k\left(1-(R_{Y/X_1X_3})^2\right)} = \frac{(10-2-1)\cdot 0{,}7709^2}{2(1-0{,}7709^2)} = 5{,}127526$$

Так как $F_{крит} < F_{набл}$, то с вероятностью 0,95 гипотеза о статистической значимости эмпирических данных принимается, корреляционная модель может быть построена.

Значение средней ошибки аппроксимации:

$$\bar{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - y_i^m}{y_i}\right|\cdot 100\% = \frac{1{,}6110057}{10}\cdot 100\% \approx 16{,}11\%$$

Это говорит о не очень высокой точности модели.

Величины средних коэффициентов эластичности:

$$K_{э_1} = a_1\frac{\overline{x_1}}{\overline{y}} = \frac{7073{,}3}{7832{,}5}\cdot 0{,}5295093 = 0{,}47 \quad K_{э_2} = a_2\frac{\overline{x_3}}{\overline{y}} = \frac{566{,}2}{7832{,}5}\cdot 1{,}69208826 = 0{,}12$$

Изменение признака X_1 на 1% влечет за собой изменения признака Y на 47, 82%, а вследствие изменения признака X_3 – на 12,23%.