День 10

Проблема мультиколлинеарности и методы его устранения

Составитель: доцент кафедры отраслевой экономики Ахмедова М.М.

План:

- 1. Понятие мультиколлинеарности.
- 2. Методы устранения или уменьшения мультиколлинеарности.

Мультиколлинеарность

Под ***мультиколлинеарностью*** *понимается высокая взаимная коррелированность объясняющих переменных.* Мультиколлинеарность может проявляться в функциональной (явной) и стохастической (скрытой) формах.

Функциональная

Стохастическая

При *функциональной форме* мультиколлинеарности по крайней мере одна из парных связей между объясняющими переменными является линейной функциональной зависимостью. В этом случае матрица $X'X$ особенная, так как содержит линейно зависимые векторы-столбцы и ее определитель равен нулю, т. е. нарушается предпосылка *6* регрессионного анализа. Это приводит к невозможности решения соответствующей системы нормальных уравнений и получения оценок параметров регрессионной модели.

Однако в экономических исследованиях мультиколлинеарность чаще проявляется в *стохастической форме*, когда между хотя бы двумя объясняющими переменными существует тесная корреляционная связь. Матрица $X'X$ в этом случае является неособенной, но ее определитель очень мал.

Функциональная мультиколлинеарность

r	x1	x2	x3	x4	x5	y
x1	1	0,5	0,6	0,7	1	0,9
x2	0,5	1	0,4	0,1	0,4	0,3
x3	0,6	0,4	1	0,3	0,2	0,2
x4	0,7	0,1	0,3	1	0,5	0,1
x5	1	0,4	0,2	0,5	1	0,8
Y	0,9	0,3	0,3	0,1	0,8	1

Стохастическая мультиколлинеарность

r	X1	x2	x3	x4	x5	y
x1	1	0,5	0,6	0,7	0,9	0,9
x2	0,5	1	0,4	0,1	0,4	0,3
x3	0,6	0,4	1	0,3	0,2	0,2
x4	0,7	0,1	0,3	1	0,5	0,1
x5	0,9	0,4	0,2	0,5	1	0,8
Y	0,9	0,3	0,3	0,1	0,8	1

Отсутсвие мультиколлинеарности

r	X1	x2	x3	x4	x5	Y
x1	1	0,5	0,6	0,7	0,4	0,9
x2	0,5	1	0,4	0,1	0,4	0,3
x3	0,6	0,4	1	0,3	0,2	0,2
x4	0,7	0,1	0,3	1	0,5	0,1
x5	0,4	0,4	0,2	0,5	1	0,8
Y	0,9	0,3	0,3	0,1	0,8	1

Между какими переменными имеется мултиколлинеарная связь?

R	x1	x2	x3	x4	x5	y
x1	1	0,5	0,6	0,7	0,4	0,9
x2	0,5	1	0,82	0,1	0,4	0,3
x3	0,6	0,82	1	0,3	0,2	0,2
x4	0,7	0,1	0,3	1	0,91	0,75
x5	0,4	0,4	0,2	0,91	1	0,8
Y	0,9	0,3	0,3	0,75	0,8	1

Интервальная оценка функции регрессии и ее параметров

Доверительный интервал для индивидуальных значений зависимой переменной. Построенная доверительная область для $M_x(Y)$ (см. рис. 3.6) определяет местоположение модельной линии регрессии (т.е. условного математического ожидания), но не отдельных возможных значений зависимой переменной, которые отклоняются от средней. Поэтому при определении *доверительного интервала для индивидуальных значений* y_0^* *зависимой переменной* необходимо учитывать еще один источник вариации — *рассеяние вокруг линии регрессии*, т.е. в оценку суммарной дисперсии $s_{\hat{y}}^2$ следует включить величину s^2. В результате оценка дисперсии индивидуальных значений y_0 при $x = x_0$ равна

Доверительный интервал для условного математического ожидания

$$\hat{y} - t_{1-\alpha;k} \cdot s_{\hat{y}} \le M_x(Y) \le \hat{y} + t_{1-\alpha;k} \cdot s_{\hat{y}}$$

$s_{\hat{y}} = \sqrt{s^2_{\hat{y}}}$ — стандартная ошибка групповой средней $\hat{y}$

$$s^2_{\hat{y}} = s^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$$

Доверительный интервал для прогнозов индивидуальных значений { y_0^* }

$$\hat{y}_0 - t_{1-\alpha;\, n-2} s_{\hat{y}_0} \le y_0^* \le \hat{y}_0 + t_{1-\alpha;n-2} s_{\hat{y}_0}$$

$$s_{\hat{y}_0}^2 = s^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right),$$

Доверительный интервал для параметров регрессионной модели. Наряду с интервальным оцениванием функции регрессии иногда представляет интерес построение доверительных интервалов для параметров регрессионной модели, в частности для параметров регрессионной модели, в частности для β_1 и σ^2.

$$t = \frac{b_1 - \beta_1}{s} \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})}^2$$

$$b_1 - t_{1-\alpha;n-2} \frac{s}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \le \beta_1 \le b_1 + t_{1-\alpha;n-2} \frac{s}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}.$$

При построении доверительного интервала для параметра σ^2 исходят из того, что статистика $\frac{ns^2}{\sigma^2}$ имеет χ^2-распределение с $k = n - 2$ *степенями свободы*. Поэтому *интервальная оценка для* σ^2 на уровне значимости α имеет вид (см. (**2.43**)):

$$\frac{ns^2}{\chi^2_{\alpha/2;n-2}} \le \sigma^2 \le \frac{ns^2}{\chi^2_{1-\alpha/2;n-2}}$$

Пример 3.3.

По данным табл. 3.1: 1) оценить сменную среднюю добычу угля на одного рабочего для шахт с мощностью пласта 8 м;

2) найти 95%-ные доверительные интервалы для индивидуального и среднего значений сменной добычи угля на 1 рабочего для таких же шахт;

3) найти с надежностью 0,95 интервальные оценки коэффициента регрессии β_1 и дисперсии σ^2.

Уравнение регрессии Y по X было получено

$$\hat{y} = -2{,}75 + 1{,}016x$$

1. Оценим условное математическое ожидание $M_{x=8}(Y)$. Выборочной оценкой $M_{x=8}(Y)$ является групповая средняя $\hat{y}_{x=8}$, которую найдем по уравнению регрессии:

$$\hat{y}_{x=8} = -2{,}75 + 1{,}016 \cdot 8 = 5{,}38 \text{ (т)}.$$

Для построения доверительного интервала для $M_{x=8}(Y)$ необходимо знать дисперсию его оценки, т.е. $s^2_{\hat{y}_{x=8}}$. Составим вспомогательную таблицу (табл. 3.2) с учетом того, что $\bar{x}=9{,}4$ (м), а значения определяются по полученному уравнению регрессии.

Расчетные формулы

$$s^2 = \frac{\sum_{i=1}^{n} \left(\hat{y}_i - y_i\right)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

$$s_{\hat{y}}^2 = s^2 \left(\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right)$$

x_i	8	11	12	9	8	8	9	9	8	12	Σ
$(x_i - \bar{x})^2$	1,96	2,56	6,76	0,16	1,96	1,96	0,16	0,16	1,96	6,76	24,40
$\hat{y}_i = -2{,}75 + {}$ $+ 1{,}016 x_i$	5,38	8,43	9,44	6,39	5,38	5,38	6,39	6,39	5,38	9,44	–
$e_i^2 = (\hat{y}_i - y_i)^2$	0,14	2,48	0,31	0,37	0,14	0,39	0,15	1,94	0,39	2,08	8,39

$$s^2 = \frac{8{,}39}{10-2} = 1{,}049 \qquad s^2_{\hat{y}_{x=8}} = 1{,}049\left[\frac{1}{10} + \frac{(8-9{,}4)^2}{24{,}4}\right] = 0{,}189$$

$$s_{\hat{y}_{x=8}} = \sqrt{0{,}189} = 0{,}435$$

Расчет доверительного интервала для математического ожидания

$$5{,}38-2{,}31\cdot 0{,}435 \le M_{x=8}(Y) \le 5{,}38+2{,}31\cdot 0{,}435,$$

$$4{,}38 \le M_{x=8}(Y) \le 6{,}38 \text{ (т)}.$$

$$t_{0{,}95;8}=2{,}31 \quad s_{\hat{y}_{x=8}}=\sqrt{0{,}189}=0{,}435$$

Итак, средняя сменная добыча угля на одного рабочего для шахт с мощностью пласта 8 м с надежностью 0,95 находится в пределах от 4,38 до 6,38 т.

Расчет доверительного интервала для индивидуальных значений

$$5{,}38-2{,}31\cdot 1{,}113\le y^*_{x_0=8}\le 5{,}38+2{,}31\cdot 1{,}113$$

$$2{,}81\le y^*_{x_0=8}\le 7{,}95.$$

$$t_{0,95;8}=2{,}31$$

$$s_{y_{x_0=8}}=\sqrt{1{,}238}=1{,}113$$

$$s^2_{y_{x_0=8}}=1{,}049\left(1+\frac{1}{10}+\frac{(8-9{,}4)^2}{24{,}4}\right)=1{,}238$$

Таким образом, индивидуальная сменная добыча угля на одного рабочего для шахт с мощностью пласта 8 м с надежностью 0,95 находится в пределах от 2,81 до 7,95 т.

Найдем 95%-ный доверительный интервал для параметра β_1.

$$1{,}016 - 2{,}31\frac{\sqrt{1{,}049}}{\sqrt{24{,}4}} \le \beta_1 \le 1{,}016 + 2{,}31\frac{\sqrt{1{,}049}}{\sqrt{24{,}4}},$$

$$0{,}537 \le \beta_1 \le 1{,}495$$

или $0{,}537 \le \beta_1 \le 1{,}495$, т. е. с надежностью 0,95 при изменении мощности пласта X на 1 м суточная выработка Y будет изменяться на величину, заключенную в интервале от 0,537 до 1,495 (т).

Найдем 95%-ный доверительный интервал для параметра σ^2.

Учитывая, что $\alpha=1-0,95=0,05$, найдем по таблице III приложений

$$\chi^2_{\alpha/2;n-2} = \chi^2_{0,025;8} = 17,53;$$

$$\chi^2_{1-\alpha/2;n-2} = \chi^2_{0,975;8} = 2,18.$$

$$\frac{10 \cdot 1,049}{17,53} \le \sigma^2 \le \frac{10 \cdot 1,049}{2,18},$$

или $\quad 0,598 \le \sigma^2 \le 4,81, \quad$ и $\quad 0,773 \le \sigma \le 2,19.$

Таким образом, с надежностью 0,95 дисперсия возмущений заключена в пределах от 0,598 до 4,81, а их стандартное отклонение — от 0,773 до 2,19 (т). ▶