3.6. ДОВЕРИТЕЛЬНЫЕ ГРАНИЦЫ ДЛЯ ВЫБОРОЧНЫХ ЗНАЧЕНИЙ КОЭФФИЦИЕНТОВ РЕГРЕССИИ

Рассмотрим простую линейную регрессионную модель зависимости y_i от x_i:

$$\hat{y}_i = a_0 + a_1 x_i , \qquad (3.6.1)$$

коэффициенты которой необходимо найти с помощью МНК.

Сделаем это, проведя центрирование переменной x_i:

$$x_i - \overline{x} . \qquad (3.6.2)$$

Тогда, подставляя в систему нормальных уравнений МНК эти значения, получим:

$$\begin{cases} \sum_{i=1}^{n} y_i = a_0 n + a_1 \sum_{i=1}^{n} (x_i - \overline{x}) \\ \sum_{i=1}^{n} y_i (x_i - \overline{x}) = a_0 \sum_{i=1}^{n} (x_i - \overline{x}) + a_1 \sum_{i=1}^{n} (x_i - \overline{x})^2 \end{cases} . \qquad (3.6.3)$$

Но поскольку выполняется очевидное условие:

$$\sum_{i=1}^{n} (x_i - \overline{x}) = 0 , \qquad (3.6.4)$$

то из первого равенства легко найти выборочное значение свободного члена уравнения регрессии:

$$a_0 = \overline{y} , \qquad (3.6.5)$$

а из второго равенства системы легко определить выборочное значение коэффициента регрессии:

$$a_1 = \frac{\sum_{i=1}^{n} y_i (x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} . \qquad (3.6.6)$$

Поскольку полученные значения коэффициентов уравнения регрессии представляют собой выборочную оценку действительных коэффициентов регрессии, необходимо оценить точность и надёжность определения этих параметров.

Обозначим истинное значение свободного члена линейной регрессии (генеральной совокупности) через A_0, а значение коэффициента линейной регрессии генеральной совокупности через A_1, и найдём - насколько выборочные значения коэффициентов отличаются от значений генеральной совокупности. Поскольку уравнение регрессии, коэффициенты которой находятся на всем множестве наблюдений (генеральной совокупности), описывают переменную y_i с некоторой ошибкой:

$$y_i = A_0 + A_1 x_i + \delta_i , \qquad (3.6.7)$$

то, просуммировав по имеющейся выборке, можно получить:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} (A_0 + A_1 x_i + \delta_i) .$$

Здесь n – размер выборки из генеральной совокупности,

δ_i – случайная нормально распределённая ошибка с нулевым математическим ожиданием. Не стоит её путать с ошибкой аппроксимации

$$\varepsilon_i = y_i - (a_0 + a_1 x_i), \qquad (3.6.8)$$

так как a_0 и a_1 – выборочные значения коэффициентов, а коэффициенты A_0 и A_1 – истинные значения генеральной совокупности.

В силу того, что выполняется (3.6.4), получим:

$$A_0 = \frac{\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \delta_i}{n} = \frac{\sum_{i=1}^{n} y_i}{n} - \frac{\sum_{i=1}^{n} \delta_i}{n}. \qquad (3.6.9)$$

Ну а поскольку ранее было показано, что a_0 определяется как средняя арифметическая (3.6.5), то полученное равенство можно представить в таком виде:

$$A_0 = a_0 + \frac{\sum_{i=1}^{n} \delta_i}{n} \rightarrow A_0 - a_0 = \frac{\sum_{i=1}^{n} \delta_i}{n}. \qquad (3.6.10)$$

Из которого следует, что отклонение выборочного значения свободного члена от его истинного значения представляет собой линейную функцию от δ_i, а поскольку последняя по условию является нормально распределённой величиной, то и разность (A_0-a_0) является нормально распределённой величиной.

В параграфе (3.4) было показано, что основанием для вычисления доверительных границ средней арифметической, как некоторой выборочной оценки математического ожидания выступает неравенство: $|\bar{x} - M(x)| < t\sigma/\sqrt{n}$, Точно также для рассматриваемого случая выполняется неравенство

$$|a_0 - A_0| < t\sigma_{a0}/\sqrt{n}. \qquad (3.6.11)$$

Заменяя дисперсию на её выборочное значение, и учитывая, что модель имеет два коэффициента, значит, число степеней свободы равно n-2, получим:

$$|a_0 - A_0| < t_\alpha \sqrt{\frac{1}{n-2} \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n}}. \qquad (3.6.12)$$

Тогда доверительные границы для этого коэффициента определяются так:

$$a_0 - t_\alpha \sqrt{\frac{1}{n-2} \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n}} < A_0 < a_0 + t_\alpha \sqrt{\frac{1}{n-2} \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n}}. \qquad (3.6.13)$$

Аналогично покажем, что коэффициент регрессии зависит от δ_i. Для этого умножим левые и правые части (3.6.7) на (3.6.2) и просуммируем полученное выражение по всем имеющимся выборочным значениям n. Получим с учётом (3.6.4):

$$\sum_{i=1}^{n} y_i (x_i - \overline{x}) = A_1 \sum_{i=1}^{n} (x_i - \overline{x})^2 + \sum_{i=1}^{n} \delta_i \,. \qquad (3.6.14)$$

Разделив теперь левую и правую стороны равенства на $\sum_{i=1}^{n} (x_i - \overline{x})^2$, и учитывая (3.6.6), получим:

$$a_1 = A_1 + \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \text{ или } a_1 - A_1 = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \,. \qquad (3.6.15)$$

Таким образом, и разность между истинным коэффициентом регрессии и его выборочным значением, найденным с помощью МНК, определяется отклонениями δ_i.

Поэтому доверительные границы для коэффициента линейной регрессии находятся аналогично по следующим формулам:

$$a_1 - t_\alpha \sqrt{\frac{1}{n-2} \frac{\sum_{i=1}^{n} \varepsilon_i^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}} < A_1 < a_1 + t_\alpha \sqrt{\frac{1}{n-2} \frac{\sum_{i=1}^{n} \varepsilon_i^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}} \,. \qquad (3.6.16)$$

Легко убедиться в том, что чем больше выборка n, тем меньше становится доверительный интервал и при $n \to \infty$ расчётные значения совпадают с истинными значениями коэффициентов генеральной совокупности.