§ 28. Элементы регрессионного анализа. Линейная регрессия.

В регрессионном анализе изучается связь между зависимой переменной Y и одной или несколькими независимыми переменными. Пусть Y зависит от одной переменной x. При этом предполагается, что Y принимает заданные (фиксированные) значения, а зависимая переменная x имеет случайный разброс из-за ошибок измерения, влияния неучтенных факторов и других причин. Каждому значению x соответствует некоторое вероятностное распределение СВ Y. Предположим, что СВ Y «в среднем» линейно зависит от значений переменной x. Это означает, что условное математическое ожидание случайной величины Y при заданном значении переменной x имеет вид

$$M(Y/x) = \beta_0 + \beta_1 x$$

Функция переменной x, определяемая правой частью этой формулы, называется *линейной регрессией* Y на x, а параметры β_0, β_1 - *параметрами линейной регрессии*. На практике параметры линейной регрессии неизвестны и их оценки определяются по результатам наблюдений переменных Y и x.

Пусть проведено n независимых наблюдений случайной величины Y при значениях переменной $x = x_1, x_2, ..., x_n$. При этом измерения величины Y дали следующие результаты: $y_1, y_2, ..., y_n$. Так как эти значения имеют разброс относительно линейной регрессии, то связь между переменными Y и x можно записать в виде *линейной* (по параметрам β_0, β_1) *регрессионной модели:*

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

где ε - случайная ошибка наблюдений, причем $M(\varepsilon) = 0$, $D(\varepsilon) = \sigma^2$. Значение дисперсии ошибок σ^2 неизвестно, и оценка ее определяется по результатам наблюдений.

Задача линейного регрессионного анализа состоит в том, чтобы по результатам наблюдений (x_i, y_i), $i = \overline{1,n}$:

1) Получить наилучшие точечные и интервальные оценки неизвестных параметров β_0, β_1 и σ^2 линейной регрессионной модели.
2) Проверить статистические гипотезы о параметрах модели.
3) Проверить, достаточно ли хорошо модель согласуется с результатами наблюдений (адекватность модели результатам наблюдений).

В соответствии с моделью результаты наблюдений зависимой переменной $Y: y_1, y_2, ..., y_n$ являются реализациями случайных величин

$$\beta_0 + \beta_1 x + \varepsilon_i,$$

обозначаемых Y_i, $i = \overline{1,n}$.

Задача линейного регрессионного анализа решается в предположении, что случайные ошибки наблюдений ε_i и ε_j не коррелированы, имеют математические ожидания. Равные нулю, и одну и ту же дисперсию, равную σ^2, т.е.

$$M(\varepsilon_i) = 0$$

$$K_{\varepsilon_i \varepsilon_j} = \begin{cases} 0, & i \neq j \\ \sigma^2, & i = j \end{cases} \qquad i = \overline{1,n}$$

При статистическом анализе регрессионной модели предполагается также, что случайные ошибки наблюдений ε_i $(i = \overline{1,n})$ имеют нормальное распределение, т.е. $\varepsilon_i \sim N(0, \sigma^2)$ $(i = \overline{1,n})$. В этом случае ошибки наблюдений также являются независимыми случайными величинами.

Для нахождения оценок параметров модели по результатам наблюдений используется *метод наименьших квадратов*. По этому методу выбирают такие оценки β_0, β_1, которые минимизируют сумму квадратов отклонений наблюдаемых значений случайных величин Y_i от их математических ожиданий, т.е.

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x))^2 .$$

МНК-оценки параметров линейной регрессии имеют вид:

$$\beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2},$$

$$\overline{\beta_0} = \bar{y} - \overline{\beta_1}\, \bar{x}.$$

Или

$$\beta_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2},$$

$$\overline{\beta_0} = \frac{\sum x^2 \sum y - \sum x \sum xy}{n\sum x^2 - (\sum x)^2}.$$

Оценки параметров линейной регрессии, вычисленные по методу наименьших квадратов, при любом законе распределения ошибок наблюдения ε_i $(i = \overline{1,n})$ при указанных выше условиях имеют следующие свойства:

1) Они являются линейными функциями результатов наблюдений y_i, $(i = \overline{1,n})$ и несмещенными оценками параметров, т.е. $M(\overline{\beta}_j) = \beta_j$, $j = 0,1$.
2) Они имеют минимальные дисперсии в классе несмещенных оценок, являющихся линейными функциями результатов наблюдений.

Если ошибки наблюдений не коррелированы и имеют нормальное распределение, то выполняется свойство:

3) МНК-оценки совпадают с оценками, вычисленными по методу максимального правдоподобия.

Функция

$$y = \overline{\beta_0} + \overline{\beta_1} x$$

определяет выборочную (эмпирическую) регрессию Y на x. Она является оценкой теоретической регрессии по результатам наблюдений. Разности между наблюдаемыми значениями переменной Y при $x = x_i$ и расчетными значениями $\tilde{y}_i = \overline{\beta_0} + \overline{\beta_1} x_i$ называются остатками и обозначаются e_i:

$$e_i = y_i - \tilde{y}_i, \ (i = \overline{1,n}).$$

Выборочное уравнение регрессии записывают еще в виде:

$$y = \overline{\beta_0} + \rho_{xy} x$$

Коэффициент ρ_{xy} называют выборочным коэффициентом регрессии Y на X.

$$\rho_{xy} = \frac{\sum n_{xy} xy - n\bar{x}\,\bar{y}}{n(\overline{x^2} - (\bar{x})^2)} = \frac{\sum n_{xy} xy - n\bar{x}\,\bar{y}}{n\tilde{\sigma}_x^2}$$

Умножим обе части равенства на $\frac{\tilde{\sigma}_x}{\tilde{\sigma}_y}$. Получим $\rho_{xy} \frac{\tilde{\sigma}_x}{\tilde{\sigma}_y} = \frac{\sum n_{xy} xy - n\bar{x}\,\bar{y}}{n\tilde{\sigma}_x \tilde{\sigma}_y}$.

Обозначим

$$r_B = \rho_{xy} \frac{\tilde{\sigma}_x}{\tilde{\sigma}_y}$$

и назовем *выборочным коэффициентом корреляции.* Здесь $\tilde{\sigma}_x, \tilde{\sigma}_y$ - выборочные с.к.о.

Таким образом,

$$r_B = \frac{\sum n_{xy} xy - n\bar{x}\,\bar{y}}{n\tilde{\sigma}_x \tilde{\sigma}_y}.$$

Выразим

$$\rho_{xy} = r_B \frac{\tilde{\sigma}_y}{\tilde{\sigma}_x}$$

тогда уравнение прямой регрессии Y на X имеет вид:

$$\bar{y}_x - \bar{y} = r_B \frac{\tilde{\sigma}_y}{\tilde{\sigma}_x}(x - \bar{x}).$$

Аналогично для регрессии X на Y:

$$\bar{x}_y - \bar{x} = r_B \frac{\tilde{\sigma}_x}{\tilde{\sigma}_y}(y - \bar{y}).$$

Замечание. Выборочный коэффициент корреляции является оценкой коэффициента корреляции

$$r = \frac{M(XY) - M(X)M(Y)}{\tilde{\sigma}_x \tilde{\sigma}_y}$$

(можно показать это методом моментов)

Гипотезу о значимости выборочного коэффициента корреляции мы рассматривали выше.

Заметим, что $r_B^2 = \rho_{xy}\rho_{yx} \Rightarrow r_B = \pm\sqrt{\rho_{xy}\rho_{yx}}$. Знак r_B совпадает со знаком ρ_{xy} и ρ_{yx}, т.к. $\rho_{yx} = r_B \frac{\tilde{\sigma}_x}{\tilde{\sigma}_y}$, $\rho_{xy} = r_B \frac{\tilde{\sigma}_y}{\tilde{\sigma}_x}$.

§ 29. Сгруппированные данные. Корреляционные таблицы.

При большом числе наблюдений одно и то же значение x может встретиться n_x раз, одно и то же значение y - n_y раз. Одна и та же пара чисел (x, y) может наблюдаться n_{xy} раз. Поэтому данные наблюдений группируют, т.е. подсчитывают частоты n_x, n_y, n_{xy}. Сгруппированные данные записывают в виде таблицы, которую называют корреляционной. Например,

Y\X	10	20	30	40	n_y
0,4	5	-	7	14	26
0,6	-	2	6	4	12
0,8	3	19	-	-	22
n_x	8	21	13	18	$n = 60$

В случае сгруппированных данных параметры регрессии вычисляются по тем же формулам, что и для несгруппированных данных.