Лекция 16

Понятие о корреляционном анализе

Вопросы лекции

1.Основные положения теории корреляции

2. Виды корреляционной взаимосвязи переменных

При ***одной независимой*** переменой X мерой ее связи с ***зависимой*** переменной Y служит ***коэффициент корреляции***.

Он рассчитывается по выборке пар наблюдений (x_i, y_i). В случае ***нескольких*** переменных последовательно вычисляются коэффициенты корреляции по нескольким рядам числовых данных. Они сводятся в таблицы, называемые ***корреляционными матрицами***.

Корреляционная матрица - квадратная матрица, на пересечении ***строк*** и ***столбцов*** которой находятся ***коэффициенты корреляции*** между соответствующими ***переменными***.

Для системы ***двух*** случайных величин

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

коэффициент корреляции рассчитывается по формуле

$$r_{XY} = \frac{(1/n) \cdot \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \cdot \sigma_y}$$

где $\bar{x}, \bar{y}$ - средние значения, а σ_x, σ_y - СКО случайных величин X, Y соответственно.

В случае ***более двух*** независимых факторов рассчитывают ***корреляционную матрицу***

$$r(i,j)=\begin{pmatrix} 1 & r_{12} & \dots & r_{1n} \\ r_{21} & 1 & \dots & r_{2n} \\ \dots & \dots & \dots & \dots \\ r_{n1} & r_{n2} & \dots & 1 \end{pmatrix},$$

симметричную относительно главной диагонали.

Пример

Имеются данные наблюдений за состоянием погоды и посещаемостью музея и парка

Число ясных дней (X_1)	Количество посетителей ***музея*** (X_2)	Количество посетителей ***парка*** (X_3)
8	495	132
14	503	348
20	380	643
25	305	865
20	348	743
15	465	541

Необходимо определить, существует ли связь между ***состоянием погоды*** и ***посещаемостью*** музеев и парков.

Расчет дает корреляционную матрицу

$$r(i,j)=\begin{pmatrix} 1 & -0{,}921 & 0{,}975 \\ -0{,}921 & 1 & -0{,}919 \\ 0{,}975 & -0{,}919 & 1 \end{pmatrix}$$

Видно, что корреляция между состоянием погоды и посещаемостью музея равна -0,921, а между состоянием погоды и посещаемостью парка 0,975. Таким образом, имеется ***отрицательная корреляция*** между ***посещаемостью музея*** и ***количеством солнечных дней*** и практически ***линейная положительная корреляция*** между ***посещаемостью парка*** и ***состоянием погоды***.

Величина ***остаточной дисперсии*** определяется отклонениями ***фактических*** данных от ***теоретических***

$$D_{ост} = \frac{1}{n}\sum (y - \hat{y}_x)^2 \quad ,$$

где y - ***фактические***, $\hat{y}_x$ - ***теоретические*** данные.

Чем ***меньше*** остаточная дисперсия, тем ***меньше*** влияние не учитываемых в уравнении регрессии факторов, и ***лучше*** его соответствие ***эмпирическим данным***.

При обработке статистических данных перебираются разные математические функции, и из них выбирается та, для которой остаточная дисперсия ***минимальна***.

В случае ***нелинейной регрессии*** вместо ***коэффициента корреляции*** используется ***индекс корреляции***:

$$p_{xy} = \sqrt{1 - \frac{\sum (y - \hat{y}_x)^2}{\sum (y - \bar{y})^2}}, \qquad 0 \le p_{xy} \le 1$$

Долю дисперсии, объясняемую ***регрессией***, характеризует ***коэффициент*** (в случае линейной регрессии) или ***индекс*** (в нелинейном случае) ***детерминации***, равный квадрату коэффициента (или индекса) корреляции.

Оценка значимости уравнения регрессии в целом

Сравнивается ***фактическое*** F_{fact} и ***табличное*** $F_{табл}$ значения критерия Фишера:

$$F_{fact} = \frac{\sum(\hat{y}-\bar{y})^2/m}{\sum(y-\hat{y})^2/(n-m-1)} = \frac{R^2}{1-R^2}\cdot\frac{(n-m-1)}{m} = \frac{r_{xy}^2}{1-r_{xy}^2}(n-2)$$

где n - число ***наблюдений***, m - число ***параметров*** при переменной x. Последнее равенство относится к случаю ***линейной*** регрессии ($m=1$).

Если $F_{fact} > F_{табл}$ то H_0 ***отклоняется*** и уравнение регрессии считается ***статистически значимым***, в противном случае H_0 ***принимается***.

Оценка значимости отдельных параметров регрессии

Ограничимся случаем парной линейной регрессии

$$y = a + b \cdot x + \varepsilon$$

По ***каждому из параметров*** определяется его ***стандартная ошибка***.

1. Стандартная ошибка ***линейного коэффициента регрессии*** b определяется по формуле:

$$m_b = \sqrt{\frac{\sum (y - \hat{y}_x)^2}{(n-2) \cdot \sum (x - \bar{x})^2}} \quad .$$

2. Стандартная ошибка ***коэффициента*** a равна:

$$m_a = \sqrt{\frac{\sum (y - \hat{y}_x)^2 \cdot \sum x^2}{(n-2) \cdot n \cdot \sum (x - \bar{x})^2}}$$

3. стандартная ошибка ***коэффициента корреляции*** –

$$m_{r_{xy}} = \sqrt{\frac{1 - r_{xy}^2}{n-2}} .$$

Рассчитывается значение t-критерия Стьюдента по формулам

$$t_b = \frac{b}{m_b}; \quad t_a = \frac{a}{m_a}; \quad t_r = \frac{r}{m_r}.$$

Выдвигается гипотеза H_0 о ***случайной*** природе показателей, т.е. о ***незначимом*** их ***отличии от нуля***.

Сравнивается ***фактическое*** и ***критическое*** (табличное) значения t - статистики при числе степеней свободы $n-2$.

Если $t_{факт} > t_{табл}$, то H_0 - ***отклоняется***, т.е. считается, что a, b, r_{xy} ***не случайно*** отличаются от нуля и сформировались под действием ***систематически действующего фактора*** x.

Если $t_{факт} < t_{табл}$, то H_0 ***принимается*** и признается ***случайная*** природа формирования одного или всех параметров регрессии.

Множественная корреляция

Практическая значимость уравнения множественной регрессии ***в целом*** оценивается с помощью ***индекса множественной корреляции***. Он характеризует меру ***совместного влияния*** факторов на результат:

$$R_{y,x_1x_2\ldots x_p} = \sqrt{1 - \frac{\sum (y - \hat{y}_{x_1x_2\ldots x_p})^2}{\sum (y - \bar{y})^2}} \quad , \quad 0 \le R_{y,x_1x_2\ldots x_p} \le 1 \quad .$$

Чем ***ближе*** R к единице, тем ***теснее*** связь результативного признака с ***набором исследуемых факторов***.

Низкое значение R означает, что в модель ***не включены существенные факторы*** или ***форма связи*** не отражает реальные соотношения между переменными.

Значимость уравнения множественной регрессии ***в целом***, как и для парной регрессии, оценивается с помощью ***F-критерия Фишера***.

Оценка значимости ***коэффициентов регрессии*** проводится по t - критерию Стьюдента.

Те коэффициенты регрессии, для которых $t_{факт} > t_{табл}$, являются ***статистически значимыми***.

Остальные ***статистически незначимы*** и формируются под воздействием ***случайных факторов***.