Множественная регрессия и корреляция

Соотношение между социально-экономическими явлениями и процессами определяются большим числом одновременно и совокупно действующих факторов.

В связи с этим часто возникает задача исследования зависимости переменной ***Y*** от нескольких объясняющих факторов: $x_1, x_2, \ldots, x_k$

Эта задача решается с помощью

множественного корреляционно-регрессионного анализа

Исходными данными для множественного анализа служит уже не два набора данных: $\{(x_i, y_i), i=1,\ldots n\}$, где ***x*** – факторный, а ***y*** – результативный признаки, а ***k+1*** набор, который можно представить в виде матрицы:

y	x_1	x_2	…	x_k
y_1	x_{11}	x_{12}	…	x_{1k}
y_2	x_{21}	x_{22}	…	x_{2k}
…	…	…	…	…
y_n	x_{n1}	x_{n2}	…	x_{nk}

Множественный корреляционно-регрессионный анализ

Задачи множественного корреляционно-регрессионного анализа:

Измерение тесноты между признаками

Отбор факторных признаков в модель

Установление неизвестных причин связей

Определение вида уравнения регрессии

Построение регрессионной модели и оценка её параметров

Проверка значимости параметров связи

Интервальное оценивание параметров связи

Требуется определить аналитическое выражение формы связи между результативным признаком ***y*** и факторными признаками $x_1, x_2, \ldots, x_k$:

$$\hat{y}_x = f(x_1, x_2, \ldots, x_k)$$

где, ***k*** – число факторных признаков

Уравнение множественной линейной регрессии

Коэффициенты уравнения регрессии, как и в случае однофакторного анализа (парной регрессии), ищутся ***методом наименьших квадратов***

Но из-за особенностей ***МНК*** в случае множественной регрессии применяются только линейные уравнения и уравнения, приводимые к линейным

Из-за трудностей обоснования формы связи чаще всего используется линейное уравнение, которое можно записать в следующей форме:

$$\hat{y}_x = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_k x_k + \varepsilon$$

Где $a_0, a_1, \ldots, a_k$ – параметры модели (коэффициенты регрессии);

ε – случайная величина (остаток).

Уравнение множественной линейной регрессии

Коэффициенты уравнения регрессии a_i показывает, на какую величину в среднем изменится результативный признак y, если переменную x_i увеличить на единицу измерения при фиксированном (постоянном) значении других факторов, входящих в уравнение регрессии.

Оценку параметров модели можно провести в матричной форме:

$$Y = X \cdot a + \varepsilon$$

где Y – вектор значений зависимой переменной $(n \times 1)$

X – матрица значений независимых переменных $x_1, x_2, \ldots, x_k$. Размерность матрицы равна $n \times (k+1)$. Первый столбец является единичным, так как в уравнении регрессии a_0 умножается на единицу.

a – подлежащий оцениванию вектор неизвестных параметров размерности $(k+1) \times 1$.

ε – вектор случайных отклонений размерности $n \times 1$

Уравнение множественной линейной регрессии

Сформулируем гипотезу модели множественной регрессии.

$$1.\ y_i = \sum_{j=1}^{k} a_i x_{ij} + \varepsilon_i,\ где\ i = 1,\ldots, n - спецификация\ модели$$

$$2.\ x_{ij} - детерминированные\ величины$$

Векторы регрессоров $\boldsymbol{x_j} = (x_{1j}, x_{2j}, \ldots, x_{nj})'$, j=1,2,…,k – линейно независимы.

(‘) – знак транспонирования.

$$3.\ E\varepsilon_i = 0,\ E\varepsilon_i^2 = V(\varepsilon_i) = \sigma^2,\ \forall i$$

$$4.\ E(\varepsilon_i \cdot \varepsilon_m) = 0$$

при $i \neq m$ – статистическая независимость (некоррелированность) ошибок для разных наблюдений.

$$5.\ \varepsilon_i \approx N(0, \sigma^2)$$

То есть ε_i – нормально распределенная случайная величина со средним значением 0 и дисперсией σ^2 (Нормальная линейная регрессионная модель)

Уравнение множественной линейной регрессии (параметры уравнения)

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1\, x_{11}\, x_{12} \dots x_{1k} \\ 1\, x_{21}\, x_{22} \dots x_{2k} \\ \dots\ \dots\ \dots\ \dots\ \dots \\ 1\, x_{n1}\, x_{n2} \dots x_{nk} \end{bmatrix}, \quad a = \begin{bmatrix} a_0 \\ a_1 \\ \dots \\ a_k \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix}$$

$$Y = X \cdot a + \varepsilon$$

$$y_1 = a_0 + a_1 x_{11} + a_2 x_{12} + \dots + a_k x_{1k} + \varepsilon_1$$
$$y_2 = a_0 + a_1 x_{21} + a_2 x_{22} + \dots + a_k x_{2k} + \varepsilon_2$$
$$\dots$$
$$y_n = a_0 + a_1 x_{n1} + a_2 x_{n2} + \dots + a_n x_{nk} + \varepsilon_n$$

Определим вектор-столбец коэффициентов ММР при помощи МНК

Уравнение множественной линейной регрессии (параметры уравнения)

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1\, x_{11}\, x_{12} \dots x_{1k} \\ 1\, x_{21}\, x_{22} \dots x_{2k} \\ \dots \dots \dots \dots \dots \\ 1\, x_{n1}\, x_{n2} \dots x_{nk} \end{bmatrix}, \quad a = \begin{bmatrix} a_0 \\ a_1 \\ \dots \\ a_k \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix} \qquad Y = X \cdot a + \varepsilon$$

$$\varepsilon = Y - \hat{Y} = Y - Xa \qquad ESS = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon \longrightarrow \min$$

$$\varepsilon'\varepsilon = (Y - Xa)'(Y - Xa) = Y'Y - \hat{Y}'Xa - a'X'Y + a'X'Xa$$

$$= Y'Y - 2aX'Y + a'X'Xa \qquad \textit{поскольку}$$

$$(Xa)' = a'X' \qquad Y'(Xa) = (Xa)'Y - \textit{скаляр}$$

$$\textit{и} \quad (a'X'X)a = ((a'X'X)a)' = a'(a'X'X)' = a'(X'Xa)$$

Уравнение множественной линейной регрессии (параметры уравнения)

$$\frac{\partial ESS}{\partial a} = -2X'Y + 2X'Xa = 0 \qquad a = (X'X)^{-1}X'Y$$

(X'X)$^{-1}$ – матрица, обратная матрице ***X'X***. Такая матрица существует в силу линейной независимости векторов $\boldsymbol{x}_j$ (п.2 гипотезы ММР).

Покажем, что вектор остатков ε ортогонален всем векторам переменных $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k$, которые являются столбцами матрицы ***X***. Данное условие ортогональности эквивалентно равенству: ***X'ε = 0***

$$X'\varepsilon = X'(Y - Xa) = X'Y - X'Xa = X'Y - X'X(X'X)^{-1}X'Y = 0$$

Используя этот факт, получим для ESS полезную формулу:

$$\varepsilon'\varepsilon = Y'Y - 2a'X'Y + a'X'Xa = Y'Y - a'(2X'Y - X'X(X'X)^{-1}X'Y) =$$

$$Y'Y - a'X'Y$$

Уравнение множественной линейной регрессии

Теорема Гаусса-Маркова.

Предположим, что:

$$Y = X \cdot a + \varepsilon$$

X – детерминированная матрица размерности ***n·(k+1)***, имеющая максимальный ранг ***k+1***.

$$E\varepsilon_i = 0,\ E\varepsilon_i^2 = V(\varepsilon_i) = \sigma^2,\ \forall i$$

Тогда МНК-оценка $a=(X'X)^{-1}X'Y$ является наиболее эффективной оценкой (обладает наименьшей дисперсией) в классе всех несмещенных оценок (Best Linear Unbiased Estimation - BLUE)

Уравнение множественной линейной регрессии (критерий Стьюдента)

Оценивание достоверности каждого из параметров модели осуществляется при помощи t-критерия Стьюдента.

Для любого из параметров модели a_j значение t-критерия рассчитывается по формуле:

$$t_{расч} = \frac{a_j}{S_\varepsilon \sqrt{b_{jj}}}$$

где S_ε – стандартное (среднее квадратическое) отклонение уравнения регрессии.

$$S_\varepsilon = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-k-1}}$$

b_{jj} – диагональные элементы матрицы $(X'X)^{-1}$

Коэффициент регрессии a_j считается достаточно надежным, если расчетное значение t-критерия Стьюдента с $(n-k-1)$ степенями свободы превышает табличное, т.е.

$t_{расч} > t_{\alpha,n-k-1}$. Если надежность не подтверждается, то следует вывод о его несущественности и устранения из модели или замены на другой факторный признак.

Уравнение множественной линейной регрессии (коэффициент эластичности)

Непосредственно с помощью коэффициентов регрессии нельзя сопоставить факторные признаки по степени их влияния на зависимую переменную из-за различия единиц измерения и разной степени колеблемости.

Для устранения таких различий применяются частные *коэффициенты эластичности* $\boldsymbol{Э_j}$ и *бета – коэффициенты* $\boldsymbol{\beta_j}$

Коэффициент эластичности: → $Э_j = a_j \frac{\bar{x}_j}{\bar{y}}$

где $\boldsymbol{a_j}$ – коэффициент регрессии фактора ***j***;

$\bar{y}$ – среднее значение результативного признака;

$\bar{x}_j$ – среднее значение признака ***j***;

Коэффициент эластичности показывает, на сколько процентов изменится зависимая переменная ***y*** при изменении фактора ***j*** на 1%

Уравнение множественной линейной регрессии (β-коэффициент)

β-коэффициент: $\Rightarrow$ $\beta_j = a_j \frac{S_{xj}}{S_y}$

где S_{xj} – среднее квадратическое отклонение фактора ***j***;
S_y – среднее квадратическое отклонение фактора ***y***

$$S_{xj} = \sqrt{\frac{\sum_{i=1`}^{n} (x_{ij} - \bar{x}_j)^2}{n}} \qquad S_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n}}$$

β-коэффициент показывает, на какую часть величины среднего квадратического отклонения S_y изменится зависимая переменная ***y*** при изменении соответствующей зависимой переменной x_j на величину своего среднего квадратического отклонения при фиксированном значении остальных независимых переменных.

Указанные коэффициенты позволяют проранжировать факторы по степени их влияния на зависимую переменную

Уравнение множественной линейной регрессии (Δ-коэффициент, R^2)

Δ-коэффициент: $\Delta_j = r_{yj} \frac{\beta_j}{R^2}$

где r_{yj} – коэффициент парной корреляции между фактором ***j*** и зависимой переменной;
R^2 – множественный коэффициент детерминации

Коэффициент множественной детерминации используют для оценки качества множественных регрессионных моделей.

Коэффициент множественной детерминации

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \varepsilon_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Коэффициент детерминации показывает долю вариации результативного признака, находящегося под воздействием факторных признаков, т.е. определяет, какая доля вариации признака ***y*** учтена в модели и обусловлена влиянием на него факторов, включенных в модель.

Чем ближе R^2 к единице, тем выше качество модели

Уравнение множественной линейной регрессии (R^2, F-критерий)

При добавлении независимых переменных значение R2 увеличивается, поэтому коэффициент R2 должен быть скорректирован с учетом числа независимых переменных по формуле:

$$R^2_{корректт} = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

Для оценки значимости модели регрессии используют F-критерий Фишера.

$$F = \frac{R^2 / k}{(1 - R^2)/(n-k-1)}$$

Если расчетные значения критерия с $\gamma_1 = k$ и $\gamma_2 = (n-k-1)$ степенями свободы больше табличного при заданном уровне значимости, то модель считается значимой.

Уравнение множественной линейной регрессии (мера точности)

В качестве меры точности модели применяют стандартную ошибку, которая представляет собой отношение суммы квадратов уровней остаточной компоненты к величине ***(n-k-1)***:

$$S_\varepsilon = \sqrt{\frac{\sum_{i=1}^{n} \varepsilon_i^2}{(n-k-1)}}$$

где $\varepsilon_i = y_i - \hat{y}_i$

Отбор факторных признаков в модель

Отбор факторов является важнейшей проблемой при построении множественных регрессионных моделей. Он проводится на основе качественного и количественного анализа социально-экономических явлений с использованием статистических и математических критериев

Проводят ***три стадии отбора факторов:***

1. Предварительное определение ***перечня факторов*** оказывающих влияние на переменную ***y***

2. Сравнительная оценка и отсев факторов

3. Окончательный выбор факторов в процессе построения разных вариантов моделей и оценки значимости их параметров

Для сравнительной оценки и отсева части факторов составляют матрицу парных коэффициентов корреляции, измеряющих тесноту линейной связи каждого фактора с результативным признаком и с каждым из остальных факторных признаков.

Матрица парных линейных коэффициентов корреляции

	y	x_1	x_2	…	x_i	…	x_n
y	1	r_{yx_1}	r_{x_2y}	…	$r_{y_ix_i}$	…	r_{yx_n}
x_1	r_{x_1y}	1	$r_{x_2x_1}$	…	$r_{x_1x_i}$	…	$r_{x_1x_n}$
x_2	r_{x_2y}	$r_{x_2x_1}$	1	…	$r_{x_2x_i}$	…	$r_{x_2x_n}$
…	…	…	…	…	…	…	…
x_i	r_{x_iy}	$r_{x_ix_1}$	$r_{x_ix_2}$	…	1	…	$r_{x_ix_n}$
…	…	…	…	…	…	…	…
x_n	r_{x_ny}	$r_{x_nx_1}$	$r_{x_nx_2}$	…	$r_{x_nx_i}$	…	1

$\boldsymbol{y}$ – результативный признак, $\boldsymbol{x_1, x_2, \ldots, x_n}$ – факторные признаки

$\boldsymbol{r_{ij}}$ – парный коэффициент корреляции между признаками $\boldsymbol{x_i}$ и $\boldsymbol{x_j}$