DOI: 10.34220/MAMSP_185-191
УДК 004

МОДЕЛИРОВАНИЕ ПРОГНОЗА ПРОДАЖ НА ОСНОВАНИИ МЕТОДОВ КОРРЕЛЯЦИИ И РЕГРЕССИИ

М.В. Чертов[1], В.И. Анциферова[1]

[1]ФГБОУ ВО «Воронежский государственный лесотехнический университет имени Г.Ф. Морозова»

В представленной статье исследованы методы корреляционного и регрессионного анализа, разработан собственный алгоритм получения исходных интересующих данных, а также представлен алгоритм расчета для закона изменения объемов продаж. Исходя из представленных вычислений, произведен расчет погрешностей коэффициентов корреляции и результатов регрессионного анализа.

Ключевые слова: корреляция, регрессия, планирование, прогноз продаж, СУБД, запросы, SQL, данные, критерий корреляции Пирсона.

SALES FORECAST MODELING BASED ON CORRELATION AND REGRESSION METHODS

M.V. Chertov[1], V.I. Antsiferova[1]

[1]Voronezh State University of Forestry and Technologies named after G.F. Morozov

The presented article explores the methods of correlation and regression analysis, developed its own algorithm for obtaining the initial data of interest, and also presents the calculation algorithm for the law of change in sales volumes. Based on the presented calculations, the errors of the correlation coefficients and the results of the regression analysis were calculated.

Keywords: correlation, regression, planning, sales forecast, DBMS, queries, SQL, data, Pearson correlation criterion.

Критерий корреляции Пирсона – это метод параметрической статистики, дающий возможность определить присутствие или отсутствие линейной связи между двумя количественными показателями, оценить ее статистическую значимость и её тесноту. Другими словами, критерий корреляции Пирсона позво-

ляет определить, имеется ли линейная связь между изменениями значений двух представленных переменных.

Критерий корреляции Пирсона позволяет определить, какова теснота (или сила) корреляционной связи двух показателей, измеренных в количественной шкале. При помощи дополнительных расчетов можно определить, насколько значимой статистически является выявленная связь [9].

Условиями и ограничениями для применения критерия Хи-квадрата Пирсона являются:

1. Сопоставляемые показатели, измеренные только в количественной шкале.

2. С помощью критерия корреляции Пирсона есть возможность определить исключительно наличие и силу линейной взаимосвязи между величинами. Другие характеристики связи, такие как направление, характер изменений, наличие зависимости одной переменной от другой - определяются уже при помощи регрессионного анализа.

3. Количество сопоставляемых величин должно быть равно двум. В ситуации, при которой необходимо провести анализ взаимосвязи трех и более параметров воспользоваться необходимо методом факторного анализа.

4. Критерий корреляции Пирсона параметрический и, учитывая данный факт, условием его применения служит нормальное распределение сопоставляемых переменных. При необходимости проведения корреляционного анализа показателей, распределение которых отлично от нормального, в том числе, показателей, измеренных в порядковой шкале, нужно использовать коэффициент ранговой корреляции Спирмена.

5. Необходимо различать и разграничивать понятия зависимости и корреляции. Зависимость величин влечет наличие корреляционной связи между ними, но не наоборот.

Наиболее распространенными методами анализа связи между количественными переменными являются методики регрессионного анализа. В эксперименте наблюдаются значения т + 1 переменных У, и X_1, X_2, ..., $X_т$.

В регрессионном анализе изучается связь между переменной Y, являющейся переменной зависимой, и переменными X_1, X_2, ..., $X_т$, являющихся независимыми. Такой вид связи описывается определенной математической моделью.

Для получения матрицы экспериментальных данных необходимо использовать один из двух способов. При первом способе значения независимых пе-

ременных X_1 и X_2, ... •, $X_т$ выбираются и устанавливаются без каких-либо погрешностей экспериментатором в каждом проводимом им опыте, при таких значениях значение зависимой переменной У измеряется с ошибками. Этот вид эксперимента будет называться активным. При данном подходе случайной величиной будет являться исключительно У. Во время использовании второго способа, будут одновременно наблюдаться значения всех взятых переменных Y, X_1, X_2, •.., $X_т$, причем все данные переменные будут случайны, т. е. матрица экспериментальных данных в таком случае будет случайной выборкой значений многомерной случайной величины (Y,X_1,X_2, ...,$X_т$). Этот вид эксперимента называется пассивным [6, 8].

Второй способ дает возможность провести корреляционный анализ, а значит, сделать статистические выводы о мерах линейной зависимости между указанными переменными.

Наиболее часто используемым методом регрессионного анализа является метод наименьших квадратов [5, 7]. Данный метод используется для получения математического описания исследуемого объекта.

По имеющимся значениям $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$

необходимо построить функцию $f(x)$.

Уравнение регрессионной зависимости выходной величины от варьируемых факторов выполняют в далее представленной последовательности:

1) выбирают вид уравнения;
2) рассчитывают коэффициенты регрессии;
3) проверяют адекватность модели.

Рассмотрим данную регрессионную модель. Предположим, что случайная величина X зависит от T, принимающей значения $t_1, t_2, ..., t_n$, при этом не являющиеся случайными.

Зависимость среднего значения случайной величины X от T будет отображаться величиной, называемой математическим ожиданием и определяться по формуле

$$M(X|T = t_i) = f(t), \quad (1)$$

где $X = f(t)$ – уравнение регрессии, $f(t)$ представляет собой кривую регрессии X на T.

Разница между математическим ожиданием и полученной в данном эксперименте величиной X

$$\varepsilon_i = X_i - M(X|T = t_i) = X_i - f(t_i), \quad (2)$$

и есть ошибка наблюдения.

Ошибки наблюдений должны удовлетворять условия Гаусса-Маркова:

1) математическое ожидание $M(\varepsilon_i) = 0$, при всех значениях i ;

2) дисперсия $D(\varepsilon_i) = \sigma^2 < +\infty$ при всех значениях i ;

3) величины независимы $cov(\varepsilon_i, \varepsilon_j) = 0$, при всех значениях i, j.

Набор $\{\varepsilon_1, \varepsilon_2, \dots \varepsilon_n\}$ является элементарными событиями одной и той же случайной величины, вследствие того, что при всех значениях $\varepsilon_i \cdot M(\varepsilon_i)$ и $D(\varepsilon_i)$ одинаковы.

Дисперсия величины ε, определяется по формуле

$$D[\varepsilon] = M[\varepsilon^2] - (M[\varepsilon])^2 = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 \quad (3)$$

или

$$D[\varepsilon] = \frac{1}{n}\sum_{i=1}^{n}(X_i - f(t_i))^2. \quad (4)$$

Допускаем, что функция $f(t)$ определяется параметрами $\vec{c} = \{c_1, c_2, \dots, c_m\}$ имеем $f(t, \vec{c})$ тогда дисперсия ошибки от неизвестных параметров равна

$$D[\varepsilon] = \frac{1}{n}\sum_{i=1}^{n}(X_i - f(t_i, \vec{c}))^2. \quad (5)$$

Дисперсия ошибок $D[\varepsilon] \subset (0,+\infty)$, имеет минимум на интервале $(0, \infty)$, который можно определить из выражения

$$\frac{\partial D[\varepsilon]}{\partial \vec{c}} = \frac{\partial D[\varepsilon]}{\partial c_0} \cdot \vec{e}_0 + \dots + \frac{\partial D[\varepsilon]}{\partial c_0} \cdot \vec{e}_m = 0, \quad (6)$$

где $\{\vec{e}_k\}_{k=0}^{m} = \{\vec{e}_0, \dots, \vec{e}_m\}$ является базисом из линейно-независимых функций

$\{\vec{e}_k\}_{k=0}^{m} = \{\varphi_k\}_{k=0}^{m} = \{\varphi_0(x), \dots, \varphi_m(x)\}$ при условии $m < n$ аппроксимирующая функция является разложением по базису: $f(x, \vec{c}) = \sum_{k=0}^{m} c_k \cdot \phi_k(x)$.

Дисперсию необходимо записать в виде:

$$D[\varepsilon] = \frac{1}{n}\sum_{i=1}^{n}(y_i - \sum_{i=o}^{m} c_k \cdot \varphi_k(x_i))^2 \quad (7)$$

Для того, чтобы найти коэффициенты регрессии продифференцируем по каждому c_j из $\{c_k\}_{k=0}^{m}$, где $j = 0,1,2,\dots,m$.

$$\begin{cases} \frac{\partial D}{\partial c_0} = 0 \\ \frac{\partial D}{\partial c_1} = 0 \\ \dots \\ \frac{\partial D}{\partial c_m} = 0 \end{cases} \quad (8)$$

$$\frac{\partial D}{\partial c_j} = \frac{\partial}{\partial c_j}\sum_{i=1}^{n}(y_i - \sum_{i=o}^{m} c_k \cdot \varphi_k(x_i))^2 = -2 \cdot \sum_{i=1}^{n}(y_i - \sum_{k=0}^{m} c_k \varphi_k(x_i)) \cdot$$

$$\sum_{k=0}^{m} \frac{\partial c_k}{\partial c_j} \varphi_k(x_i) = 0 \quad (9)$$

$$\frac{\partial c_k}{\partial c_j} = \begin{cases} 1, j = k \\ 0, j \neq k \end{cases} \equiv \delta_{jk} \quad (10)$$

Зная, что $\sum_{k=0}^{m} \delta_{jk} \varphi_k(x_i) = \phi_j(x)$, имеем

$$\sum_{i=1}^{n}(y_i - \sum_{k=0}^{m} c_k \varphi_k(x_i)) \cdot \phi_j(x_i) = 0, \quad (11)$$

где $j = 0,1,2,\dots,m$.

Полученное выражение можем представить в виде системы уравнений:

$$\begin{cases} \sum_{i=1}^{n}(y_i - \sum_{k=0}^{m} c_k \varphi_k(x_i)) \cdot \phi_0(x_i) = 0 \\ \sum_{i=1}^{n}(y_i - \sum_{k=0}^{m} c_k \varphi_k(x_i)) \cdot \phi_1(x_i) = 0 \\ \dots \\ \sum_{i=1}^{n}(y_i - \sum_{k=0}^{m} c_k \varphi_k(x_i)) \cdot \phi_m(x_i) = 0 \end{cases} \quad (12)$$

В матричном виде система уравнений имеет вид: $A \cdot \vec{c} = \vec{b}$, где

$$A = \begin{pmatrix} (\phi_0, \phi_0)(\phi_1, \phi_0)\dots(\phi_m, \phi_0) \\ (\phi_0, \phi_1)(\phi_1, \phi_1)\dots(\phi_m, \phi_1) \\ \dots \\ (\phi_0, \phi_m)(\phi_1, \phi_m)\dots(\phi_m, \phi_m) \end{pmatrix}, \quad \vec{c} = \begin{pmatrix} c_0 \\ c_1 \\ \dots \\ c_m \end{pmatrix}, \quad \vec{b} = \begin{pmatrix} (y, \phi_0) \\ (y, \phi_1) \\ \dots \\ (y, \phi_m) \end{pmatrix} \quad (13)$$

Решая систему, получим вектор коэффициентов аппроксимирующей функции:

$$\vec{c} = \{c_0, c_1, \dots c_m\} = A^{-1} \cdot b =$$

$$= \begin{pmatrix} (\phi_0, \phi_0)(\phi_1, \phi_0)\dots(\phi_m, \phi_0) \\ (\phi_0, \phi_1)(\phi_1, \phi_1)\dots(\phi_m, \phi_1) \\ \dots \\ (\phi_0, \phi_m)(\phi_1, \phi_m)\dots(\phi_m, \phi_m) \end{pmatrix}^{-1} \begin{pmatrix} (y, \phi_0) \\ (y, \phi_1) \\ \dots \\ (y, \phi_m) \end{pmatrix} \quad (14)$$

Аппроксимирующая функция в итоге будет иметь вид:

$$f(x) = c_0\phi_0(x) + c_1\phi_1(x) + \dots + c_m\phi_m(x) \quad (15)$$

При выборе степенной функции аппроксимирующая отражается в виде:

$$f(x, \vec{c}) = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + \dots + c_m x^m = \sum_{k=0}^{m} c_k x^k \quad (16)$$

Система $A \cdot \vec{c} = \vec{b}$ принимает вид:

$$A = \begin{pmatrix} \sum x_1 & \sum x_i & \dots \sum x_i^m \\ \sum x_i & \sum x_i^2 & \dots \sum x_i^{m+1} \\ \dots & & \\ \sum xm_i & \sum x_i^{m+1} & \dots \sum x_i^{2m} \end{pmatrix}, \quad \vec{b} = \begin{pmatrix} \sum y_i \\ \sum y_i x_i \\ \dots \\ \sum y_i x_i^m \end{pmatrix} \qquad (17)$$

$$\text{где } \vec{c} = \begin{pmatrix} c_0 \\ c_1 \\ \dots \\ c_m \end{pmatrix} = A^{-1}b = \begin{pmatrix} \sum x_1 & \sum x_i & \dots \sum x_i^m \\ \sum x_i & \sum x_i^2 & \dots \sum x_i^{m+1} \\ \dots & & \\ \sum xm_i & \sum x_i^{m+1} & \dots \sum x_i^{2m} \end{pmatrix}^{-1} \begin{pmatrix} \sum y_i \\ \sum y_i x_i \\ \dots \\ \sum y_i x_i^m \end{pmatrix} \qquad (18)$$

$$f(x) = \begin{pmatrix} c_0 \\ c_1 \\ \dots \\ c_m \end{pmatrix} \{1, x, \dots x^m\} = f(x, \vec{c}) = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + \dots + c_m x^m \qquad (19)$$

Модель такого вида применять нельзя, если истинная зависимость отклика $y = f(x)$ от некоторого фактора:

1) будет иметь более одного экстремума;

2) имеет точку перегиба;

3) при некоторых значениях действующих факторов функция изменяется скачком.

При возникновении первых двух случаев рекомендуется выбирать многочлены более высокого порядка.

Список литературы

1. Антамошкина, О.И. Имитационные модели прогноза продаж / О.И. Антамошкина, Ю.В. Булгаков, О.В. Зинина // Вестник КРАСГАУ. – 2011. – № 2 (53). – С. 28-36.

2. Бахтеева, Р.Х. Статистический анализ себестоимости от продаж и оказанных услуг / Р.Х. Бахтеева // Синергия наук. – 2018. – № 20. – С. 203-210.

3. Глухова, Л.В. Оптимизация планирования производства на основе моделирования прогноза объемов продаж / Л.В. Глухова, И.А. Маштаков // Вестник Волжского университета им. В.Н.Татищева. – 2012. – № 1 (25). – С. 110-117.

4. Юров, А.Н. Проектирование автоматизированной системы производственных планировок / А.Н. Юров // Моделирование систем и процессов. – 2019. – Т. 12, № 1. – С. 87-93.

5. Построение интеллектуальных систем управления информационными процессами в условиях неопределенности / Ю.Ю. Громов, В.Е. Дидрих, И.В.

Дидрих, А.Ю. Гречушкина // Моделирование систем и процессов. – 2018. – Т. 11, № 1. – С. 10-14.

6. Оксюта, О.В. Разработка математической модели оптимального функционирования транспортно-логистического комплекса / О.В. Оксюта, В.А. Коротких // Моделирование систем и процессов. –2017. –Т. 10, № 3. – С. 55-66.

7. Савченко, А.Л. Анализ формирования регрессионных моделей удельных потерь энергии для различных типов тяжелых заряженных частиц / А.Л. Савченко //Моделирование систем и процессов. –2019. –Т. 12, № 2. –С. 85-93.

8. Лавлинский, В.В. Математические зависимости формализации процедур проектирования МОП-транзисторов / В.В. Лавлинский, А.Л. Савченко, А.Ю. Кулай // Моделирование систем и процессов. – 2018. – Т. 11, № 1. – С. 31-38.

9. Юров, А.Н. Проектирование автоматизированной системы производственных планировок / А.Н. Юров // Моделирование систем и процессов. – 2019. – Т. 12, № 1. – С. 87-93.