11. Линейный и нелинейный многомерный регрессионный анализ

Одним из методов, используемых для прогнозирования, является регрессионный анализ.

Регрессия **-** это статистический метод, который позволяет найти уравнение, наилучшим образом описывающее совокупность данных.

Линейный многомерный регрессионный анализ

В общем виде многомерная линейная регрессионная модель зависимости y от объясняющих переменных x_1, x_2,…, x_k имеет вид: $\tilde{y} = M(y/x_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$

.Для оценки неизвестных параметров β_j взята случайная выборка объема n из (k+1)–мерной случайной величины (y, x_1, x_2,…, x_k).

В матричной форме модель имеет вид:

$Y = X\beta + \varepsilon$,

где $Y = \begin{pmatrix} y_1 \\ y_2 \\ \ldots \\ \ldots \\ y_n \end{pmatrix}$, $X = \begin{pmatrix} 1\, x_{11}\, x_{12} \ldots x_{1k} \\ 1\, x_{21}\, x_{22} \ldots x_{2k} \\ \ldots\ldots\ldots\ldots \\ 1\, x_{n1}\, x_{n2} \ldots x_{nk} \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_k \end{pmatrix}$, $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \varepsilon_n \end{pmatrix}$

- вектор-столбец фактических значений зависимой переменной размерности n;

- матрица значений объясняющих переменных размерности n*(k+1);

- вектор-столбец неизвестных параметров, подлежащих оценке, размерности (k+1);

- вектор-столбец случайных ошибок размерности n с математическим ожиданием ME=0 и ковариационной матрицей $V(\varepsilon) = M(\varepsilon\, \varepsilon^T) = \sigma^2 E_n$ соответственно, при этом

$E_n = \begin{pmatrix} 1\,0\,0 \ldots 0 \\ 0\,1\,0 \ldots 0 \\ \ldots\ldots\ldots\ldots \\ 0\,0\,0 \ldots 1 \end{pmatrix}$ -единичная матрица размерности (nxn).

Оценки неизвестных параметров β_j находятся методом наименьших квадратов, минимизируя скалярную сумму квадратов $Q = (Y - X\beta)^T (Y - X\beta)$ по компонентам вектора β.

Далее подставив выражение

$(Y - X\beta) = \begin{pmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{pmatrix} - \begin{pmatrix} \beta_0 + \sum_{j=1}^{k} x_{1j}\beta_j \\ \beta_0 + \sum_{j=1}^{k} x_{2j}\beta_j \\ \ldots\ldots\ldots\ldots \\ \beta_0 + \sum_{j=1}^{k} x_{nj}\beta_j \end{pmatrix} = \begin{pmatrix} y_1 - \beta_0 - \sum_{j=1}^{k} x_{1j}\beta_j \\ y_2 - \beta_0 - \sum_{j=1}^{k} x_{2j}\beta_j \\ \ldots\ldots\ldots\ldots \\ y_n - \beta_0 - \sum_{j=1}^{k} x_{2j}\beta_j \end{pmatrix}$ в

$Q = (Y - X\beta)^T (Y - X\beta)$,

получаем скалярную сумму квадратов $Q = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{k} x_{ij}\beta_j)^2$

Условием обращения полученной суммы в минимум является система нормальных уравнений:

$\frac{\partial Q}{\partial \beta_j} = 0$, (j=0,1,2,…,k) .

В результате дифференцирования получается:

$2X^T(Y - X\beta) = 0$.

При замене вектора неизвестных параметров β на оценки, полученные методом наименьших квадратов, получаем следующее выражение: $X^T Y = X^T X b$.

Далее умножив обе части уравнения слева на матрицу $(X^T X)^{-1}$, получим

$(X^T X)^{-1} \cdot (X^T Y) = (X^T X)^{-1} \cdot (X^T X) b$

Так как $(X^T X)^{-1}(X^T X) = E$, тогда $b = (X^T X)^{-1}(X^T Y)$.

Полученные оценки вектора b являются не смещенными и эффективными.

Ковариационная матрица вектора b имеет вид:

$V(b) = \sigma^2 (X^T X)^{-1}$, где σ^2 - остаточная дисперсия.

Элементы главной диагонали этой матрицы представляют собой дисперсии вектора оценок b. Остальные элементы являются значениями коэффициентов ковариации:

$\mathrm{cov}(b_i b_j) = M(b_i - \beta_i)(b_j - \beta_j)$, где $i = 1 \div n$, $j = 0 \div k$.

Таким образом, оценка b_j - это линейная функция от зависимой переменной. Она имеет нормальное распределение с математическим ожиданием β_j и дисперсией $D_{b_i} = \sigma^2 \&*[(X^T X)^{-1}]_{jj}$.

Несмещенная оценка остаточной дисперсии определяется по формуле:

$\hat{S}^2_{ост} = \frac{1}{n-k-1}(Y - Xb)^T (Y - Xb)$, где n – объем выборочной совокупности; k – число объясняющих переменных.

Для проверки значимости уравнения регрессии используют F-критерий дисперсионного анализа, основанного на разложении общей суммы квадратов отклонений на составляющие части:

$Q_{общ} = Q_R + Q_{ост}$, где $Q_R = (Xb)^T (Xb) = \sum_{i=1}^{n} \hat{Y}_i^2$ - сумма квадратов отклонений (от нуля), обусловленная регрессией;

$$Q_{ост} = (Y - Xb)^T (Y - Xb) = \sum_{i=1}^{n} e_i^2$$ - сумма квадратов отклонений фактических значений зависимой переменной от расчетных $\hat{Y} = Xb$, т.е. сумма квадратов отклонений относительно плоскости регрессии, обусловленное воздействием случайных и неучтенных в модели факторов.

Для проверки гипотезы $H_0 : \beta = 0$ используется величина $F_H = \frac{\frac{1}{k+1} Q_R}{\frac{1}{n-k-1} Q_{ост}}$, которая имеет F-распределение Фишера с числом степеней свободы $\nu_1 = k+1$ и $\nu_2 = n-k-1$. Если $F_H > F_{кр}$, то уравнение регрессии значимо, т.е. в уравнении есть хотя бы один коэффициент регрессии, отличный от нуля.

В случае значимости уравнения регрессии проверяется значимость отдельных коэффициентов регрессии. Для проверки нулевой гипотезы $H_0 : \beta_j = 0$ используется величина $F_H = \frac{b_j^2}{\hat{S}^2 [(X^T X)^{-1}]_{jj}}$, которая имеет F-распределение Фишера с числом степеней свободы $\nu_1 = 1$ и $\nu_2 = n-k-1$; $[(X^T X)^{-1}]_{jj}$ - соответствующий элемент главной диагонали ковариационной матрицы.

Коэффициент регрессии β_j считается значимым, если $F_H > F_{кр}$. Для значимых коэффициентов регрессии можно построить доверительные интервалы, используя формулу $\beta_j \in \{b_j \pm t_\gamma \hat{S}[(X^T X)^{-1}]_{jj}\}$, где t_γ находится по таблице

распределения Стьюдента для уровня значимости $\alpha = 1 - \gamma$ и числа степеней свободы $\nu = n - k - 1$.

Коэффициент детерминации

Чем меньше разброс значений остатков около линии регрессии по отношению к общему разбросу значений, тем, очевидно, лучше прогноз. Например, если связь между переменными X и Y отсутствует, то отношение остаточной изменчивости переменной Y к исходной дисперсии равно 1.0. Если X и Y жестко связаны, то остаточная изменчивость отсутствует, и отношение дисперсий будет равно 0.0. В большинстве случаев отношение будет лежать где-то между этими экстремальными значениями, т.е. между 0.0 и 1.0. 1.0 минус это отношение называется R-квадратом или коэффициентом детерминации. Это значение непосредственно интерпретируется следующим образом. Если имеется R-квадрат равный 0.4, то изменчивость значений переменной Y около линии регрессии составляет 1-0.4 от исходной дисперсии; другими словами, 40% от исходной изменчивости могут быть объяснены, а 60% остаточной изменчивости остаются необъясненными. В идеале желательно иметь объяснение если не для всей, то хотя бы для большей части исходной изменчивости. Значение R-квадрата является индикатором степени подгонки модели к данным (значение R-квадрата близкое к 1.0 показывает, что модель объясняет почти всю изменчивость соответствующих переменных).

Интерпретация коэффициента множественной корреляции R.

Обычно, степень зависимости двух или более предикторов (независимых переменных или переменных X) с зависимой переменной (Y) выражается с помощью коэффициента множественной корреляции R. По определению он равен корню квадратному из коэффициента детерминации. Это неотрицательная величина, принимающая значения между 0 и 1. Для интерпретации направления связи между переменными смотрят на знаки (плюс или минус) регрессионных коэффициентов или B-коэффициентов. Если B-коэффициент положителен, то связь этой переменной с зависимой переменной положительна (например, чем больше IQ, тем выше средний показатель успеваемости оценки); если B-коэффициент отрицателен, то и связь носит отрицательный характер (например, чем меньше число учащихся в классе, тем выше средние оценки по тестам). Конечно, если B-коэффициент равен 0, связь между переменными отсутствует.

Коэффициент множественной детерминации характеризует, на сколько процентов построенная модель регрессии объясняет вариацию значений результативной переменной относительно своего среднего уровня, т. е. показывает долю общей дисперсии результативной переменной, объяснённой вариацией факторных переменных, включённых в модель регрессии.

Коэффициент множественной детерминации также называется количественной характеристикой объяснённой построенной моделью регрессии дисперсии результативной переменной. Чем больше значение коэффициента множественной детерминации, тем лучше построенная модель регрессии характеризует взаимосвязь между переменными.

Для коэффициента множественной детерминации всегда выполняется неравенство вида:

$$R^2(y, x_1 \dots x_{n-1}) \le R^2(y, x_1 \dots x_n),$$

Следовательно, включение в линейную модель регрессии дополнительной факторной переменной xn не снижает значения коэффициента множественной детерминации.

Коэффициент множественной детерминации может быть определён не только как квадрат множественного коэффициента корреляции, но и с помощью теоремы о разложении сумм квадратов по формуле:

$$R^2(y, x_1 \dots x_n) = 1 - \frac{ESS}{TSS},$$

где ESS (Error Sum Square) – сумма квадратов остатков модели множественной регрессии с n независимыми переменными:

$$ESS = \sum_{i=1}^{n} (y_i - \tilde{y}(y, x_1 \dots x_n))^2 ;$$

TSS (TotalSumSquare) – общая сумма квадратов модели множественной регрессии с n независимыми переменными:

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 .$$

Однако классический коэффициент множественной детерминации не всегда способен определить влияние на качество модели регрессии дополнительной факторной переменной. Поэтому наряду с обычным коэффициентом рассчитывают также и скорректированный (adjusted) коэффициент множественной детерминации, в котором учитывается количество факторных переменных, включённых в модель регрессии:

$$R^2_{Adj} = 1 - \frac{\frac{ESS}{(n-h)}}{\frac{TSS}{(n-1)}} = 1 - (1 - R^2) \cdot \frac{(n-1)}{(n-h)},$$

где n – количество наблюдений в выборочной совокупности;
h – число параметров, включённых в модель регрессии.
При большом объёме выборочной совокупности значения обычного и скорректированного коэффициентов множественной детерминации отличаться практически не будут.

Нелинейный многомерный регрессионный анализ

Когда данные указывают на то, что функция их распределения не совсем линейна, мы можем свести ее к линейным формам. Например, ниже приведенные нелинейные формы приводятся к линейным:

Степенная: $Y = b_0 \cdot X_1^{b_1} \cdot \dots \cdot X_n^{b_n}$ => прологарифмируем левую и правую части уравнения =>

$$Ln(Y) = Ln(b_0) + b_1 \cdot Ln(X_1) + \dots + b_n \cdot Ln(X_n)$$

Логарифмическая:

$Y = b_0 + b_1 \cdot Ln(X_1) + \dots + b_n \cdot Ln(X_n)$ => заменим выражения $Ln(X_1),\dots,Ln(X_n)$ на $U_1,\dots,U_n$ =>

$$Y = b_0 + b_1 \cdot U_1 + \dots + b_n \cdot U_n$$

Экспоненциальная: $Y = b_0 \cdot e^{b_1 X_1} \cdot \dots \cdot e^{b_n X_n}$ =>

прологарифмируем левую и правую части уравнения =>

$$Ln(Y) = Ln(b_0) + b_1 X_1 + \dots + b_n X_n$$

Полиноминальная (многочлен):

$Y = b_0 + b_1 X_1^1 + b_2 X_1^2 + \dots + b_n X_1^n$ => заменим выражения X_1^1, X_1^2, X_1^n на $U_1,\dots,U_n$ =>

$Y = b_0 + b_1 U_1 + b_2 U_2 + \dots + b_n U_n$ ***И использовать ту из них, которая лучше описывает связь между зависимой Y и независимыми переменными X_i.***

При этом следует не забывать, что в дальнейшем необходимо осуществить обратное преобразование всех коэффициентов $b_0, b_1, \dots, b_n$.