1. Регрессионный анализ. Основы

Основные задачи регрессионного анализа:

а) подбираем класс функций для анализа;

б) производим отбор наиболее информативных переменных;

в) вычисляем оценки значений параметров модели;

г) анализируем точность уравнения связи и его параметров;

д) анализируем степень пригодности уравнения для целей прогноза.

1. Регрессионный анализ. Основы

Основные варианты статистических взаимосвязей между переменными *X* и *Y*:

- Не направленная связь, обе переменные равноценны, наличие и сила взаимосвязи – *корреляционный анализ*.
- выделяем одну величину как *независимую* (*объясняющую*, *факторную*, *предиктор*), а другую как *зависимую* (*объясняемую*, *результирующую*), изменение первой может служить причиной для изменения второй переменной в виде какой либо **зависимости** – *регрессионный анализ*

Количество, качество и неоднозначность связи.

1. Регрессионный анализ. Основы

Анализ влияния объясняющей переменной на зависимую переменную «**в среднем**»:

$$M(Y\,|\,x) = f(x)$$

функция регрессии Y на X. Здесь X – объясняющая переменная (*регрессор*), Y – зависимая.

Общий вид регрессии, суть:

$$M(Y\,|\,x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n)$$

Парная, множественная регрессия. Ф. Далтон 19 в.

Возможность прогноза.

1. Регрессионный анализ. Основы

Случайность зависимости – не совпадение реальных значений зависимой переменной с её *условным математическим ожиданием*.

Дополнение случайным слагаемым ε, отражающим это несоответствие

$$Y = M(Y \mid x) + \varepsilon$$

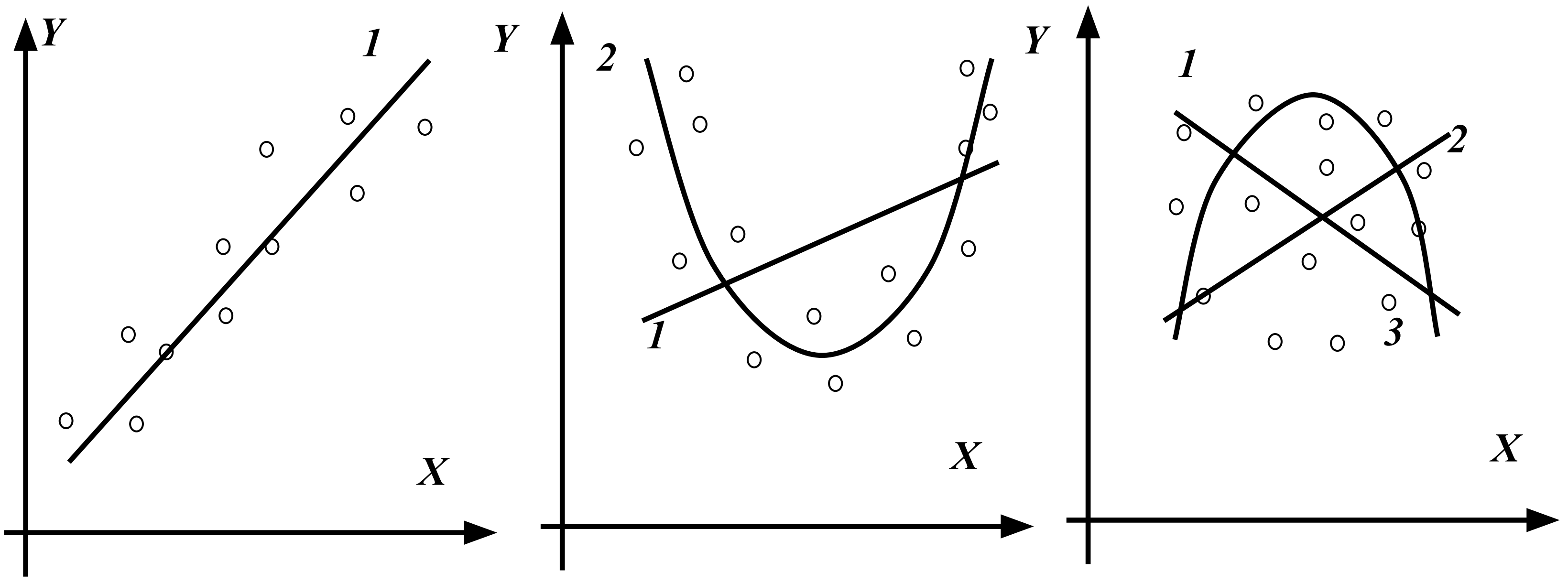
$$Y = M(Y \mid x_1, x_2, \ldots, x_n) + \varepsilon$$

Регрессионные модели (*уравнения*). Необходимость случайного члена.

1. Регрессионный анализ

1. Выбор вида связи переменных - *спецификация* уравнения *регрессии*.

Графический метод по *корреляционному полю* (*диаграмме рассеивания*). Аналитический.



1. Регрессионный анализ

2. Определение параметров (коэффициентов) модели - *параметризация*.

3. Проверка качества *регрессии* - *верификация*.

Линейная модель как общая тенденция процесса.

Параметризация линейной модели:

-Метод средних (Метод Маркова, на основе МЗР);

-Метод максимального правдоподобия.

Частный случай ММП при квадратичной целевой функции – метод наименьших квадратов МНК.

1. Регрессионный анализ

Задачи *линейного регрессионного анализа* для статистических данных (x_i, y_i) для переменных X и Y:

а) получить наилучшие по какому-либо критерию оценки неизвестных теоретических параметров a_i ;

б) проверить статистические гипотезы о параметрах модели;

в) проверить адекватность модели данных наблюдений;

г) выполнить оценку точности результатов вычислени

д) провести анализ полученных данных.

1. Парный регрессионный анализ

Задача: подобрать для 2 рядов X и Y оптимальную модель из класса линейных вида

$$y_i = ax_i + b + e_i$$

чтобы наилучшим образом линейно выразить Y через X. С учетом коррекции имеем

$$\hat{y}_i = y_i + v_i = ax_i + b$$

Решение на основе:

- метода наименьших квадратов в натуральных величинах
- метода наименьших квадратов в центрированных величинах
- Байесовского метода (на основе характеристик многомерного закона распределения)

1. Парный регрессионный анализ

1. Для решения модели по МНК в натуральных величинах:

-на поправки *v* накладывают условие метода наименьших квадратов: найти такие значения коэффициентов *a* и *b* чтобы сумма квадратов поправок v – $[v^2] = v^T v \rightarrow \min$ – была минимальной из всех возможных наборов пар коэффициентов. Расчетные формулы:

- выражаем поправки *v*, записывая уравнения поправок для модели,

$$v_i = \hat{y}_i - y_i = ax_i + b - y_i$$

или в матричном виде

$$v = \hat{y} - y = Xk - y$$

Здесь матрица плана X и искомых коэффициентов имеет вид

$$X = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}$$

$$k = \begin{pmatrix} a \\ b \end{pmatrix}$$

1. Парный регрессионный анализ

- составляем целевую функцию качества Ф для МНК

$$\Phi = \left[v^2 \right] = v^T v = \left[(ax + b - y)^2 \right]$$

которая должна быть минимальна

-для минимизации берем частные производные от Ф по k и полученное выражение приравниваем к нулю

$$\frac{\partial \Phi}{\partial k} = \frac{\partial \Phi}{\partial v} \cdot \frac{\partial \mathrm{v}}{\partial k} = 2\mathrm{v}^{\mathrm{T}} \cdot X = 0$$

с учетом того, что $v = Xk - y$

Транспонируя выражение имеем

$$X^{\mathrm{T}} \cdot \mathrm{v} = 0$$

- лемма Гаусса.

1. Парный регрессионный анализ

Подставим в лемму значение поправки $v = Xk - y$

Получаем окончательно

$$X^{\mathrm{T}} \cdot (X \cdot k - y) = X^{\mathrm{T}} X \cdot k - X^{\mathrm{T}} y = 0$$

матричную систему нормальных уравнений (совместная, 2 уравнения, 2 неизвестных) в свернутом виде

$$Nk = d,$$

с элементами

$$N = X^{\mathrm{T}} X = \begin{bmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{bmatrix} = \begin{bmatrix} [x^2] & [x] \\ [x] & n \end{bmatrix}, \quad k = \begin{pmatrix} a \\ b \end{pmatrix}, \quad d = X^{\mathrm{T}} y = \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = \begin{pmatrix} [xy] \\ [y] \end{pmatrix}$$

1. Парный регрессионный анализ

Систему в матричном виде целесообразно решать через обращение матрицы N, откуда имеем

$$k = \begin{pmatrix} a \\ b \end{pmatrix} = N^{-1} \cdot d = Q \cdot d$$

Здесь матрица $Q = N^{-1}$ носит название обратная матрица.

Модельные значения в матричном виде будут

$$\hat{y} = XQX^T \cdot y = X \cdot k$$

Поправки в измерения получим как

$$v = \hat{y} - y = (XQX^T - E)y$$

2. Решение по МНК в центрированных величинах.

Для этого исходная модель

$$\hat{y}_i = y_i + v_i = ax_i + b$$

центрируется средним

$$\hat{y} - \overline{\hat{y}} = a\left(x - \overline{x}\right) \rightarrow \hat{y} = y + v = \overline{y} + a\left(x - \overline{x}\right)$$

т.к. $\hat{y} - \overline{\hat{y}} = \hat{y} - \overline{y}$

Имеем 1 неизвестное *a*. Тогда целевая целевая функция Ф будет

$$\Phi = \left[\mathrm{v}^2\right] = \left[\left(\overline{y} + a\left(x - \overline{x}\right) - y\right)^2\right] = \left[\left(a \cdot v_x - v_y\right)^2\right]$$

Минимизация Ф дает

$$\frac{\partial \Phi}{\partial a} = \frac{\partial \Phi}{\partial v} \cdot \frac{\partial v}{\partial a} = 2\left[\left(av_x - v_y\right) \cdot v_x\right] = 0$$

1. Парный регрессионный анализ

Откуда

-нормальное уравнение

$$\left[v_x^2\right]\cdot a=\left[v_x\cdot v_y\right]$$

- решение относительно коэффициента a

$$a=\frac{\left[v_x\cdot v_y\right]}{\left[v_x^2\right]}$$

Для вычисления величины сдвига b раскроем скобки в центрированной модели модели и перегруппируем что получится

$$\hat{y}=a\cdot x+\left(\overline{y}-a\cdot\overline{x}\right)=a\cdot x+b$$

Откуда величина сдвига b будет

$$b=\overline{y}-a\cdot\overline{x}$$

1. Парный регрессионный анализ

Разделим числитель и знаменатель выражения для *a*

$$a = \frac{\left[v_x \cdot v_y\right]}{\left[v_x^2\right]}$$

на число элементов в ряде *n*. Получим т.о. его оценку в терминах элементов ковариационной матрицы

$$a = \frac{\operatorname{cov}(x,y)}{D(x)} = r_{xy} \cdot \frac{\sigma_y}{\sigma_x}$$

Подставив вид для *a* в центрированную модель, получим

$$\hat{y} = \bar{y} + \frac{\operatorname{cov}(x,y)}{D(x)} \cdot (x - \bar{x}) =$$

$$= \bar{y} + r_{xy} \cdot \frac{\hat{\sigma}_y}{\hat{\sigma}_x}(x - \bar{x})$$

1. Парный регрессионный анализ

Полученное уравнение - оценка условного математического ожидания для нормально распределенной пары рядов. Метод получения уравнения регрессии в таком виде – метод средних, метод Маркова, метод условного математического ожидания, байесовский метод.

Этот же результат можно получить на основе выборочной ковариационной матрицы для 2 рядов x и y:

$$K = \begin{pmatrix} \dfrac{\left[v_x^2\right]}{n} & \dfrac{\left[v_x v_y\right]}{n} \\ \dfrac{\left[v_y v_x\right]}{n} & \dfrac{\left[v_y^2\right]}{n} \end{pmatrix} = \begin{pmatrix} D(x) & \operatorname{cov}(x,y) \\ \operatorname{cov}(y,x) & D(y) \end{pmatrix} = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix}$$

Тогда имеем:

1. Парный регрессионный анализ

$$a = \frac{\operatorname{cov}(y,x)}{D(x)} = \frac{K_{21}}{K_{11}}$$

$$b = (\overline{y} - a \cdot \overline{x}) = \overline{y} - \frac{K_{21}}{K_{11}} \cdot \overline{x}$$

$$\hat{y} = \overline{y} + \frac{\operatorname{cov}(x,y)}{D(x)} \cdot (x - \overline{x}) =$$

$$= \overline{y} + K_{21} \cdot K_{11}^{-1} \cdot (x - \overline{x})$$

-т.е. все необходимые элементы модели в терминах ковариационной матрицы (2 мерный случай байесовского метода)

1. Парный регрессионный анализ

Если известна обратная ковариационная матрица $K^{-1} = \Lambda$ формулы для расчета коэффициентов примут вид

$$a = -\frac{\Lambda_{12}}{K_{22}}$$

и

$$b = \left(\overline{y} - a \cdot \overline{x}\right) = \overline{y} + \frac{\Lambda_{12}}{\Lambda_{22}} \cdot \overline{x}$$

Часто бывают более удобные чем другие.

В конце расчетов всегда записывается вид модели в числах.

1. Парный регрессионный анализ

Оценка точности уравнения регрессии:

-оценка модели в целом;

- оценка коэффициентов модели.

При необходимости – оценка значимости коэффициентов.

Чаще выполняется на основе метода наименьших квадратов и теоремы о переносе ошибок.

Для оценки точности модели:

-величины поправок v есть оценки истинных погрешностей Δ модели;

- обычная формула стандартного отклонения для качества модели

$$\hat{\sigma}_0 = \sqrt{\frac{v^{\mathrm{T}} v}{n-t}} = \sqrt{\frac{\Phi}{r}}$$

1. Парный регрессионный анализ

Получение погрешности модели на основе метода условного математического ожидания по условной дисперсии вида

$$\hat{\sigma}^2_{y|x} = \hat{\sigma}^2_0 = \left(K_{22} - K_{21} \cdot K_{11}^{-1} \cdot K_{12}\right) \cdot \left(\frac{n}{n-t}\right) = \left(K_{22} - \frac{K_{12}^2}{K_{11}}\right) \cdot \left(\frac{n}{n-t}\right) =$$

$$= D(y) \cdot \left(1 - r_{xy}^2\right) \cdot \left(\frac{n}{n-t}\right)$$

если используется ковариационная матрица и

$$\hat{\sigma}^2_0 = \frac{1}{\Lambda_{22}} \cdot \left(\frac{n}{n-t}\right)$$

Если использовать обратную ковариационную матрицу Λ

1. Парный регрессионный анализ

Погрешности определения коэффициентов получим на основе теоремы о переносе ошибок используя формулу где вектор коэффициентов k линейно выражен через вектор измерений y:

$$k = QX^{\mathrm{T}} \cdot y = F \cdot y$$

Тогда вид ковариационной матрицы K_k для вектора коэффициентов k

$$K_k = F \cdot K_y \cdot F^{\mathrm{T}} = \left(QX^{\mathrm{T}}\right) \cdot K_y \cdot \left(XQ\right)$$

Вектор измерений y – один – $K_y = \hat{\sigma}_0^2$ и

$$K_k = F \cdot K_y \cdot F^{\mathrm{T}} = \hat{\sigma}_0^2 \cdot Q$$

размера (2×2) - диагональные элементы – дисперсии вычисленных коэффициентов a и b.

1. Парный регрессионный анализ

Некоторые дополнительным возможностям регрессионного анализа:

-построение регрессии по методу наименьших квадратов для полиномиальной модели,

- представление на основе МНК периодических данных.

Очевидно, что это далеко не все дополнительные возможности парного регрессионного анализа.

Пусть появилось подозрение что исходные данные лучше будут представлены не линейной, а квадратичной моделью вида

$$\hat{y} = y + \mathrm{v} = a \cdot x^2 + b \cdot x + c$$

или в матричном виде

$$y + \mathrm{v} = X \cdot k$$

1. Парный регрессионный анализ

Целевая функция Ф для модели при использовании МНК будет

$$\Phi = \left[\mathrm{v}^2 \right] = \left[\left(a \cdot x^2 + b \cdot x + c - y \right)^2 \right]$$

производные от Ф по коэффициентам *a*, *b* и *c* приравненные к нулю

$$\frac{\partial \Phi}{\partial a} = \frac{\partial \left[\left(a \cdot x^2 + b \cdot x + c - y \right)^2 \right]}{\partial a} = 2 \left[\left(a \cdot x^2 + b \cdot x + c - y \right) \cdot x^2 \right] = 0$$

и т.д., откуда первое нормальное уравнение

$$\left[x^4 \right] \cdot a + \left[x^3 \right] \cdot b + \left[x^2 \right] \cdot c = \left[x^2 y \right]$$

и т.д., которых будет 3 по числу коэффициентов модели. Система совместная и разрешима относительно коэффициентов.

1. Парный регрессионный анализ

Рациональнее в матричном виде. Уравнения поправок

$$\mathrm{v} = X \cdot k - y$$

где матрица плана X и свободный член y имеют вид

,

$$X = \begin{pmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ \vdots & \vdots & \vdots \\ x_n^2 & x_n & 1 \end{pmatrix} \qquad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

целевая функция $\Phi = \mathrm{v}^T \mathrm{v}$, а её минимизация (производная по k и приравнивание к нулю дает лемму Гаусса

$$\frac{\partial \Phi}{\partial k} = \frac{\partial \Phi}{\partial v} \cdot \frac{\partial \mathrm{v}}{\partial k} = 2\mathrm{v}^{\mathrm{T}} \cdot X = 0$$

1. Парный регрессионный анализ

Транспонируя выражение имеем более привычный вид

$$X^{\mathrm{T}} \cdot v = 0$$

Подстановка в неё вида уравнений поправок v дает совместную систему нормальных уравнений

$$X^{\mathrm{T}} \cdot (X \cdot k - y) = X^{\mathrm{T}} X \cdot k - X^{\mathrm{T}} y = 0$$

или

$$Nk = d,$$

с элементами

$$N = X^{\mathrm{T}} X = \begin{pmatrix} N_{11} & N_{12} & N_{13} \\ N_{21} & N_{22} & N_{23} \\ N_{31} & N_{32} & N_{33} \end{pmatrix}, \quad k = \begin{pmatrix} a \\ b \\ c \end{pmatrix}, \quad d = X^{\mathrm{T}} y = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix}$$

1. Парный регрессионный анализ

Подход может расширяться на любую степень полинома, изменяется только состав матрицы X.

Оценка точности стандартная для МНК:

-погрешность модели,

-погрешности коэффициентов,

-если необходимо, статистический анализ.

По погрешности модели можно сказать о необходимости усложнения модели.

1. Парный регрессионный анализ

Модель процесса может быть разная . Очень распространена периодическая функция представления данных.

При построении регрессии (аппроксимации) для периодической функции модель имеет следующий вид

$$\hat{y} = k_0 + k_1 \cdot \sin(x) + k_2 \cdot \cos(x) + k_3 \cdot \sin(2x) + k_4 \cdot \cos(2x) + \ldots$$

Модели 1, 2 и т.д. порядков. Гармоники.

Для получения коэффициентов модели 1 порядка на основе МНК запишем уравнения поправок, целевую функцию Ф и минимизируем её.

- уравнения поправок

$$\mathrm{v} = k_0 + k_1 \cdot \sin(x) + k_2 \cdot \cos(x) - y$$

1. Парный регрессионный анализ

- целевая функция Φ

$$\Phi=\left[v^2\right]=\left[\left(k_0+k_1\cdot\sin\left(x\right)+k_2\cdot\cos\left(x\right)-y\right)^2\right]$$

-Минимизация

$$\begin{cases}\dfrac{\partial\Phi}{\partial k_0}=2\cdot\left[\left(k_0+k_1\cdot\sin(x)+k_2\cdot\cos(x)-y\right)\cdot 1\right]=0\\ \dfrac{\partial\Phi}{\partial k_1}=2\cdot\left[\left(k_0+k_1\cdot\sin(x)+k_2\cdot\cos(x)-y\right)\cdot\sin(x)\right]=0\\ \text{и т.д.}\end{cases}$$

и окончательно

$$\begin{cases}n\cdot k_0+\left[\sin(x)\right]\cdot k_1+\left[\cos(x)\right]\cdot k_2=\left[y\right]\\ \left[\sin(x)\right]\cdot k_0+\left[\sin^2(x)\right]\cdot k_1+\left[\cos(x)\cdot\sin(x)\right]\cdot k_2=\left[y\cdot\sin(x)\right]\\ \ldots\text{и т.д.}\end{cases}$$

1. Парный регрессионный анализ

Периодические функции имеют циклы и на основе свойств циклических перестановок имеем

$$[\sin(x)] = [\cos(x)] = [\cos(x)\cdot\sin(x)] = 0,$$

$$\left[\sin^2(x)\right] = \left[\cos^2(x)\right] = \frac{n}{2}$$

Откуда

$$\begin{cases} n\cdot k_0 + 0\cdot k_1 + 0\cdot k_2 = [y] \\ 0\cdot k_0 + \left[\sin^2(x)\right]\cdot k_1 + 0\cdot k_2 = [y\cdot\sin(x)] \\ \ldots\text{и т.д.} \end{cases}$$

и явное решение

$$\begin{cases} k_0 = \dfrac{[y]}{n} \\ k_1 = \dfrac{2\cdot[y\cdot\sin(x)]}{n} \\ k_2 = \dfrac{2\cdot[y\cdot\cos(x)]}{n} \end{cases}$$

1. Парный регрессионный анализ

Матрица системы нормальных уравнений имеет диагональный вид для модели любого порядка. Тогда по аналогии новые коэффициенты будут

$$\begin{cases} k_3 = \dfrac{2 \cdot \left[y \cdot \sin(2x) \right]}{n} \\ k_4 = \dfrac{2 \cdot \left[y \cdot \cos(2x) \right]}{n} \\ \text{и т.д.} \end{cases}$$

а вычисленные ранее **не меняются.** Таким образом, модель 2 порядка получается из модели 1 порядка простым добавлением новых коэффициентов, 3 – добавление в модель 2 порядка вычисленных коэффициентов 3 порядка и т.д.

Оценка точности стандартная. Погрешность модели, погрешность коэффициентов, если надо – статистический анализ.

1. Парный регрессионный анализ

Контрольные вопросы по теории 3 модуля

1. Основы общего регрессионного анализа. Основные этапы.
2. Парный линейный регрессионный анализ. Основные положения.
3. Решение задачи регрессии на основе МНК
4. Решение задачи регрессии на основе условного математического ожидания (метод средних).
5. Оценка точности и элементы статистического анализа в парной линейной регрессии.
6. Расширение парной модели регрессии на другие случаи.