МАТЕМАТИЧЕСКАЯ СТАТИСТИКА
Лекция 13
Регрессионный анализ-2

Повтор

§ 11. Элементы линейного регрессионного анализа

Потребность прогнозировать одни величины по значениям других

Круг задач, связанных с построением (восстановлением) зависимостей между группами числовых переменных

$$x \equiv (x_1, x_2 \dots x_p) \text{ и } y = (y_1, y_2 \dots y_m).$$

Предполагается, что

x — независимые переменные (факторы)

y — зависимых переменных (откликов). По имеющимся эмпирическим данным

$$(x^i, y^i),\ i = 1, 2, \dots, n$$

требуется построить функцию

$$y \approx f(x) = ?.$$

чтобы в дальнейшем прогнозировать

$$\hat{y} = f(x) \text{ по значению } x.$$

Будем считать, что

$$y = f(x) + \delta, \tag{1}$$

где $f(x)$ — закономерное изменение y от x;

δ — случайная составляющая с нулевым средним,

$$M(y \mid x) = f(x)$$

и называется *регрессией y по x* (рис. 15). Случайная составляющая δ:

- внутренняя изменчивость y,
- влияние факторов, неучтенных в x,
- и то, и другое вместе.

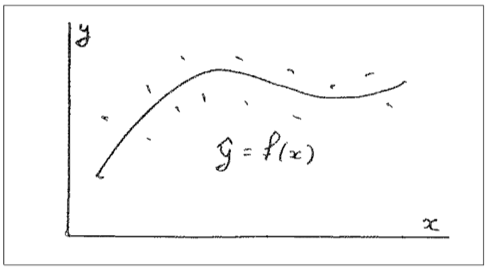


Рис. 15. Подбор функции *f* по наблюдениям

Рассматриваем неслучайные значения факторов, однако, при случайных значениях формулы получаются такие же.

Множество ***допустимых функций* - *параметрическое***:

$$\mathbf{f(x) = f(x, b)},$$

где **b** — неизвестный параметр.

Если функция $f(x, \theta)$ линейна по θ:

$$f(x, \theta) = h(x) \cdot \theta,$$

то регрессионный анализ называется линейным.

Множественная регрессия. Схема Гаусса-Маркова

А. Анализируемая модель.

$f(x)$ линейна по x, $x = (x_1, x_2 \dots x_k)$- значения k факторов и
 η — результат наблюдения, скалярная случайная величина**:**

$$\eta = f(x) + \delta = b_1 x_1 + b_2 x_2 + \dots + b_k x_k + \delta, \qquad (1)$$

$b_1, b_2 \dots b_k$ — неизвестные коэффициенты регрессии.

Пусть n раз измерены значения факторов и соответствующие значения отклика η. Результаты n измерений:

$$\eta_1 = x_{11} b_1 + \dots + x_{k1} b_k + \delta_1,$$
$$\dots\dots\dots \qquad (1а)$$
$$\eta_n = x_{1n} b_1 + \dots + x_{kn} b_k + \delta_n$$

(**первый индекс - номер фактора**, а **второй —номер наблюдения**). Предполагается также, что

$\delta_1, \delta_2 \dots \delta_n$ — некоррелированные случайные величины с одинаковыми дисперсиями

$$M(\delta_i \delta_j) = 0, \ i \neq j, \text{ и } M\delta_i = 0, \ \ M\delta_i^2 = \sigma^2. \qquad (1б)$$

Примечание [MA1]: Верная запись? Ошибки нет в нижних индексах?
Добавлен квадрат (Ю.Г.)

Соотношения (1а) удобно записывать в матричной форме:

$$\eta = X^T b + \delta, \qquad (1в)$$

$\eta = (\eta_1, \eta_2 \dots \eta_n)^T$ — вектор-столбец зависимой переменной; T — символ транспонирования;

$b = (b_1, b_2 \dots b_k)^T$ — вектор-столбец (размерности k) неизвестных коэффициентов регрессии;

$\delta = (\delta_1, \delta_2 \dots \delta_n)^T$ — вектор-столбец случайных отклонений;

$$X = \begin{bmatrix} x_{11} & \dots & x_{1n} \\ \vdots & \dots & \vdots \\ x_{k1} & \dots & x_{kn} \end{bmatrix}$$

— матрица $k \times n$ значений факторов; в ***j*-м столбце находятся значения факторов при *j*-м наблюдении**.

Б. Оценка коэффициентов регрессии. Пусть β является оценкой для b, тогда

$$\hat{\eta} = X^T \beta$$

Примечание [MA2]: Снять выделение жирным

- вектор прогнозов для имеющихся значений откликов η.

Построим оценку β для вектора b так, чтобы вектор прогнозов $\hat{\eta} = X^T\beta$ минимально отличался от вектора η измеренных значений :

Примечание [MA3]: Снять выделение жирным

$$S(\beta) \equiv \|\eta - \hat{\eta}\|^2 = (\eta - X^T\beta)^T (\eta - X^T\beta) \xrightarrow{\quad} \min \text{ по } \beta. \qquad (2)$$

Примечание [MA4]: Снять выделение жирным

После дифференцирования получили систему

$$\left(\frac{dS(\beta)}{d\beta}\right)^T = X(\eta - X^T\beta) = 0. \qquad (3)$$

Примечание [MA5]: Снять выделение жирным

$$XX^T\beta = X\eta$$

Если $\det XX^T \neq 0$, то

$$\beta = (XX^T)^{-1}X\eta = Z^{-1}X\eta, \qquad (4)$$

где $Z = XX^T$ - обозначение.

Формула для $S(.)$. при отклонении от β. Пусть $\beta_1 = \beta + \varepsilon$.

$$S(\beta + \varepsilon) = S(\beta) + \varepsilon^T Z\varepsilon \geq S(\beta) \qquad (5)$$

Свойства полученной оценки (4):

1. **Оценка линейна** по наблюдениям η.
2. **Оценка несмещённо** оценивает b. Действительно, $\eta = X^T b + \delta$

$$M\beta = b. \qquad (6)$$

3. Определим дисперсионную матрицу $D\beta$ оценки. Учитывая (4), (6) и (1в), имеем

$$(\beta - b) = Z^{-1}X\delta.$$

Тогда ковариационная матрица для β есть

$$D\beta = M(\beta - b)(\beta - b)^T = \sigma^2 Z^{-1}. \qquad Z^{-1} = [z^{ij}] \qquad (7)$$

$$\mathrm{cov}(\beta_i, \beta_j) = \sigma^2 z^{ij}$$

4. **Свойство оптимальности** оценки — **теорема Гаусса-Маркова.** *В классе линейных несмещенных оценок в условиях* (1б) *оценки* β_i, $i = 1, 2 \ldots k$ (4) *являются оценками с минимальной дисперсией*.

Конец повтора

В. Оценка дисперсии σ^2 ***наблюдений***. Наша функция

$$S(\beta) \equiv \|\eta - \hat{\eta}\|^2 = (\eta - X^T\beta)^T(\eta - X^T\beta)$$

Рассмотрим значение функции $S(\beta)$ в точке $\beta = b$- истин. знач:

$$S(b) = (\eta - X^T b)^T(\eta - X^T b) = \delta^T\delta = \sum_{i=1}^{n}\delta_i^2;$$

здесь учтено, что $\eta = X^T b + \delta$.

Математическое ожидание $MS(b)$:

$$MS(b) = n\sigma^2. \qquad (8)$$

Определим $MS(b)$ иначе. запишем $S(b)$ с учетом (5) :

$$S(b) = S(\beta + (b - \beta)) = S(\beta) + (b - \beta)^T Z(b - \beta).$$

$$MS(b) = MS(\beta) + M(b - \beta)^T Z(b - \beta). \qquad (9)$$

Определим математическое ожидание второго слагаемого в (9) в виде суммы с учетом (7): $D\beta = \sigma^2 Z^{-1}$. (т.е. $\mathrm{cov}(\beta_i, \beta_j) = \sigma^2 z^{ij}$) будет иметь вид:

$$M(b - \beta)^T Z(b - \beta) = \sum_{i=1}^{k}\sum_{j=1}^{k} z_{ij}\,\mathrm{cov}(\beta_i, \beta_j) = \sigma^2 \sum_{i=1}^{k}\sum_{j=1}^{k} z_{ij} z^{ij} =$$

$$= \sigma^2 \mathrm{tr}(ZZ^{-1}) = \sigma^2 \mathrm{tr}(E_k) = k\sigma^2,$$

здесь обозначены элементы симметричной матрицы Z через z_{ij}, а элементы матрицы Z^{-1} — через z^{ij}. $Z = [z_{ij}]$, $Z^{-1} = [z^{ij}]$

Примечание [MA6]: E_k — что обозначает? Не нужно ли курсивом?

обозначено E_k- единичная матрица размерности k. Итак,

$$MS(b) = MS(\beta) + k\sigma^2$$

Приравнивая математическое ожидание $S(b)$ выражения (9) значению $n\sigma^2$ из (8), имеем

$$\mathrm{M}S(b) = \mathrm{M}S(\beta) + k\sigma^2 = n\sigma^2,$$

откуда получаем

$$\mathrm{M}\frac{S(\beta)}{n-k} = \sigma^2.$$

Это равенство означает, что несмещенной оценкой для σ^2 является

$$s^2 = \frac{S(\beta)}{n-k} = \frac{\|\eta - \hat{\eta}\|^2}{n-k} = \frac{\|\eta - \mathrm{X}^{\mathrm{T}}\beta\|^2}{n-k}. \qquad (10)$$

Вектор

$$e \equiv \eta - \hat{\eta} = \eta - \mathrm{X}^{\mathrm{T}}\beta$$

называется остаточным вектором или вектором невязок.

Г. Доверительные интервалы для коэффициентов регрессии.

$$\beta = (\mathrm{X}\mathrm{X}^{\mathrm{T}})^{-1}\mathrm{X}\eta = \mathrm{Z}^{-1}\mathrm{X}\eta,$$

Для построения доверительного интервала нужно знать закон распределения оценки. Заметим, что каждый элемент оценки-это сумма некоррелированных случайных величин. Можно предположить, что распределение суммы приближенно нормально.

Так и сделаем: предположим, что случайные составляющие δ_j, j = 1, 2…n, распределены нормально $N(0, \sigma^2)$, тогда оценки распределены нормально:

$$\beta_i, \sim N(b_i, \sigma^2 z^{ii}).\ i = 1, 2\ldots k,$$

Кроме того, известна теорема, (она обобщает теорему о совместном распределении выборочных среднего и дисперсии при нормальных наблюдениях),

Напоминание: Теорема о совместном распределении выборочных среднего и дисперсии для нормальной совокупности:

$$\xi = (\xi_1, \xi_2\ldots\xi_n) \sim N(m, \sigma^2), \qquad \hat{m} = \bar{\xi} = \frac{1}{n}\sum_{i=1}^{n}\xi_i, \qquad s^2 = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - \hat{m})^2.$$

s^2 - сумма квадратов разностей между наблюдениями и оценками их мат. ожиданиями

1) статистики $\bar{\xi}$ и s^2 независимы;

2) с.в. $\sqrt{n}(\bar{\xi} - m)/\sigma \sim N(0, 1)$;

3) с.в. $ns^2/\sigma^2 \sim \chi^2(n-1)$.

Примечание [MA7]: E_k — что обозначает? Не нужно ли курсивом?

Примечание [MA8]: Снять выделение жирным

Примечание [MA9]: Снять выделение

Примечание [MA10]: Верхний коэффициент точно такой — ii?

Обобщение которой является **утверждение (теорема):**

$$\|\eta - \hat{\eta}\|^2 = \|\eta - X^T\beta\|^2 = \qquad S(\beta)/\sigma^2 = \|\eta - \hat{\eta}\|^2 / \sigma^2 \sim \chi^2_{n-k}$$

распределено по закону хи-квадрат χ^2_{n-k} с $(n-k)$ степенями свободы, и β и $S(\beta)$ независимы. Тогда: ясно, что

$$\frac{(\beta_i - b_i)}{\sigma\sqrt{z^{ii}}} \sim N(0,1), \qquad \frac{S(\beta)}{\sigma^2(n-k)} \sim \frac{\chi^2_{n-k}}{n-k}$$

$\beta_i, \sim N(b_i, \sigma^2 z^{ii})$.

Напоминание *Распределение Стьюдента***.** Многие задачи статистики приводят к рассмотрению следующей случайной величины.

Пусть с.в.: $\alpha, \sim N(0,1)$,

и с.в. χ^2_k, ~ хи-квадрат с k степенями свободы.

Образуем новую случайную величину T_k следующим образом:

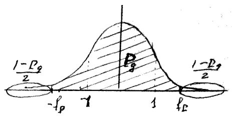
$$T_k = \alpha / \sqrt{\chi^2_k / k}.$$

Распределение этой случайной величины называется ***распределением Стьюдента***

Составляем отношение:

$$\frac{(\beta_i - b_i)}{\sigma\sqrt{z^{ii}}} \Big/ \sqrt{\frac{S(\beta)}{\sigma^2(n-k)}} = \frac{\beta_i - b_i}{s\sqrt{z^{ii}}} \equiv T_{n-k} \qquad (11)$$

неизвестный параметр σ сокращается, а $\dfrac{S(\beta)}{n-k} = s^2$. Оно подчиняется закону



Стьюдента с $(n-k)$ степенями свободы. Соответственно, доверительный интервал определяется соотношением

$$\left|\frac{\beta_i - b_i}{s\sqrt{z^{ii}}}\right| < T_P \qquad |\beta_i - b_i| < T_P s\sqrt{z^{ii}}, \qquad (12)$$

где T_P =Q((1+$P_д$)/2, $(n-k)$) — квантиль уровня (1 +$P_д$)/2 распределения Стьюдента с $(n-k)$ степенями свободы; $P_д$ — коэффициент доверия.

Проверка гипотезы о незначимости коэф-та регрессии $b_i = 0$

Основываясь на статистике (11), проверяется гипотеза $b_i = 0$ о нулевом значении коэффициента регрессии ; если $b_i = 0$, то $\dfrac{\beta_i}{s\sqrt{z^{ii}}} \sim T_{n-k}$ и тогда если

$$\left|\beta_i / s\sqrt{z^{ii}}\right| > T_P, \qquad (13)$$

то гипотеза о нулевом значении отклоняется, влияние i–го фактора есть.

Д. Нелинейная функция регрессии f(x). Связь между факторами x и откликом η может быть нелинейной, например, в виде полинома:

$$\eta = P(x) + \delta,$$

где $P(x) = b_1 + b_2 x + \ldots + b_k x^{k-1}$, $(k-1)$ — степень полинома;

$b_1, b_2 \ldots b_k$ — неизвестные коэффициенты;

δ — случайная составляющая;

$$M\delta = 0, D\delta = \sigma^2.$$

Примечание [MA11]: Снять выделение

Примечание [MA12]: Верхний коэффициент точно такой — ii?

Пусть n раз измерены значения фактора x и соответствующие значения отклика η: $(x_1, \eta_1), (x_2, \eta_2),\ldots, (x_n, \eta_n)$,.

Результаты n измерений:

$$\eta_i = b_1 + b_2 x_j + \ldots + b_k x_j^{k-1} + \delta_j, \quad j = 1, 2\ldots n, \qquad (14)$$

или, как и (1в), в матричной форме:

$$\eta = X^T b + \delta, \qquad (15)$$

$$\text{где } X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \\ & \cdots & & \\ x_1^{k-1} & x_2^{k-1} & \cdots & x_n^{k-1} \end{bmatrix}.$$

Имеем задачу (1в), и потому все формулы (2) — (13) оказываются справедливыми. Слово «линейный» в названии «линейный регрессионный анализ» означает линейность относительно неизвестных коэффициентов b_i, но не относительно факторов x_i.

Кроме полиномиальной широко используются

Другие нелинейные модели связи отклика с факторами

1) логарифмическая — если зависимость

$$\eta = a_1 x^{a_2},$$

(с увеличением фактора x значение отклика монотонно увеличивается); после логарифмирования получаем

$$\ln \eta = \ln a_1 + a_2 \ln x = b_1 + b_2 \ln x;$$

2) гиперболическая (при обратной зависимости, т.е. при увеличении x, признак η уменьшается)

$$\eta = b_1 + \frac{b_2}{x};$$

3) тригонометрическая

$$y = b_1 + b_2 \sin\omega x + b_3 \cos \omega x$$

и другие.

E. Обобщение: нелинейная функция регрессии.

Пусть связь между p факторами $(x_1, x_2\ldots x_p) \equiv x$ и откликом η выражается следующим образом:

$$\eta = f(x,b) + \delta = \sum_{i=1}^{k} b_i \varphi_i(\mathbf{x}) + \delta,$$

где $\varphi_1(x)$, $\varphi_2(x)$, , …,$\varphi_k(x)$, (— система некоторых функций. Имеется n наблюдений при различных значениях $x \equiv (x_1, x_2\ldots x_p)$:

$$(x^1, \eta_1), (x^2, \eta_2)\ldots (x^n, \eta_n):$$

$$\eta_j = \sum_{i=1}^{k} b_i \varphi_i(\mathbf{x}^j) + \delta_j, \quad j = 1, 2\ldots n,$$

или в матричной форме:

$$\eta = X^T b + \delta,$$

где X — матрица $k \times n$, в j-м столбце которой находятся элементы $\varphi_1(x^j)$, $\varphi_2(x^j)$ ... $\varphi_k(x^j)$; обозначения η, b, δ аналогичны (1в).

$$X = \begin{bmatrix} \varphi_1(x_1) & \varphi_1(x_2) & \cdots \varphi_1(x_n) \\ \varphi_2(x_1) & \varphi_2(x_2) & \cdots \varphi_2(x_n) \\ & \cdots & \\ \varphi_k(x_1) & \varphi_k(x_2) & \cdots \varphi_k(x_n) \end{bmatrix}$$

Получили задачу (1в), и потому все формулы (4) — (13) оказываются справедливыми.

Пример: отклик η нужно приблизить многочленом второй степени от двух переменных x и y, т.е.

$$f(x,y,b) = b_0 + b_1x + b_2y + b_3x^2 + b_4xy + b_5y^2,$$

сформируем наши факторы: 1, x, y, x^2, xy. y^2.

в матрице X в j-м столбце будут находиться следующие 6 элементов:

$$1, x_j, y_j, x_j^2, x_jy_j, y_j^2.$$