Регрессионный анализ данных

Оценки модели линейной регрессии. Построение модели

(1) $Y_i = \eta(\mathbf{x}_i) + \varepsilon_i$, $i = 1, \ldots, n$,

где $\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_q]'$ - q-мерная *неслучайная векторная переменная*

ε_i – случайная ошибка

$$\eta(\mathbf{x}, \boldsymbol{\beta}) = \beta_0 + \beta_1 \mathbf{x}_1 + \ldots + \beta_q \mathbf{x}_q.$$

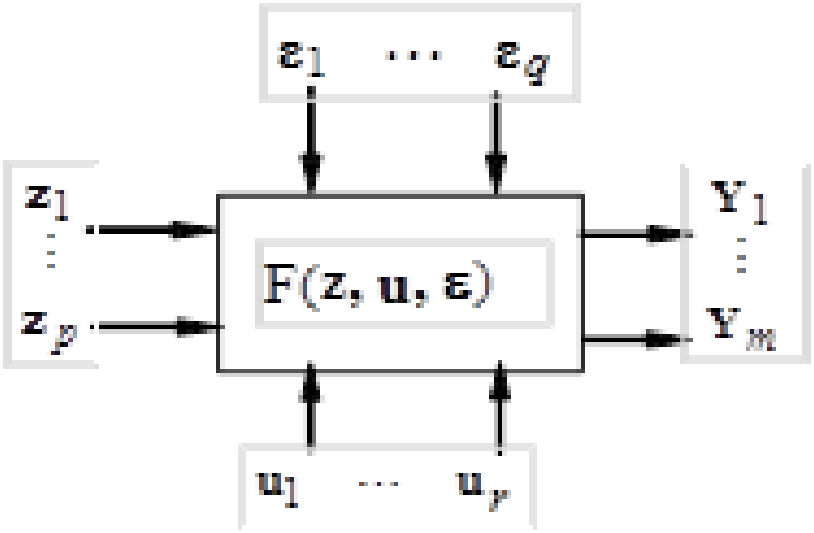
где $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_q]'$ – вектор неизвестных параметров (коэффициентов)

(2) $$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{x}_1 + \ldots + \beta_q \mathbf{x}_q + \boldsymbol{\varepsilon}$$

Если в (2.2) $\mathbf{M}\boldsymbol{\varepsilon} = 0$

(3) $$\mathbf{M}[\mathbf{Y} / \mathbf{x}] = \beta_0 + \beta_1 \mathbf{x}_1 + \ldots + \beta_q \mathbf{x}_q$$

Пример 1



1. Контролируемые неуправляемые переменные z = (z1, …, zp),
2. Контролируемые управляющие переменные u = (u1, …, ur),
3. Неконтролируемые неуправляемые переменные $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_q)$, ,
4. Контролируемые управляемые переменные Y = (Y1, …, Ym),

Линейная регрессионная модель

$\eta(x,\beta)$ - нелинейная относительно вектора параметров β

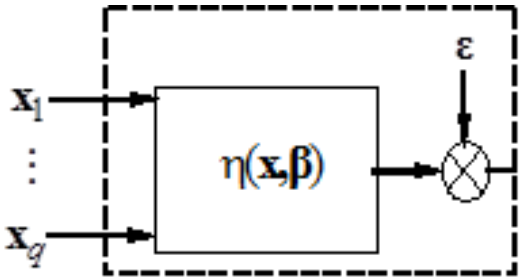


Рис. 2. Структурная схема объекта

$$Y_i = \eta(0,\ldots,0) + \sum_{j=1}^{q}\left(\frac{\partial\eta}{\partial x_j}\right)_0 x_{ij} + \frac{1}{2}\sum_j\sum_k\left(\frac{\partial^2\eta}{\partial x_j \partial x_k}\right)_0 x_{ij}x_{ik} + \ldots + \varepsilon_i$$

Обозначив постоянные

$$\beta_0 = (0,\ldots,0) \quad \beta_j = (\partial\eta/\partial x_j)_0 \qquad \beta_{jk} = (1/2)(\partial^2\eta/\partial x_j \partial x_k)_0$$

получим

$$Y_i = \beta_0 + \sum_{j=1}^{q}\beta_j x_{ij} + \sum_j\sum_k \beta_{jk} x_{ij}x_{ik} + \ldots + \varepsilon_i$$

полиномиальная модель q-го порядка одной переменной

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_q x_i^q + \varepsilon_i,$$

Нелинейная модель

$$Y_i = \beta_0 + \beta_1 e^{-\beta_2 x_i} + \varepsilon_i$$

Оценивание параметров

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{10} & x_{11} & \cdots & x_{1g} \\ x_{20} & x_{21} & \cdots & x_{2g} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{ng} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_g \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

(4) $$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

Метод наименьших квадратов (МНК)

(least squares method)

Необходимо по наблюдениям $(x_{i1}, \ldots, x_{ig}, Y_i)$, $i = 1, \ldots, n$ **найти наилучшую оценку вектора** $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_g)'$ **уравнения регрессии** $\eta(x,\beta)$.

Метод наименьших квадратов

1. **Число регрессоров q полинома (3) априори известно точно.**
2. **Все регрессоры измеряются без ошибок, а вычисления проводятся абсолютно точно.**
3. **Остаток ε является независимой нормально распределенной случайной величиной с нулевым средним $M[\varepsilon] = 0$ и неизвестной постоянной дисперсией σ_ε^2 при всех *i* = 1, …, *n*.**
4. **Дисперсия отклика *Yi* постоянна, или является известной функцией номера наблюдения *i* = 1, …, *n*.**
5. **Распределение *Yi* одинаково при всех i = 1, …, *n*.**
6. **Число опытов *n* существенно больше числа регрессоров q.**

Если условия 1÷6 соблюдаются и X′X обратимая матрица, то согласно фундаментальной теореме Гаусса-Маркова наилучшей оценкой вектора коэффициентов β является оценка $\hat{\beta}$, доставляющая минимум суммы квадратов остатков (невязок, ошибок, помех):

$$Q(\hat{\beta}) = \sum_{i=1}^{n} \left(Y_i - \sum_{k=0}^{q} x_{ik} \hat{\beta}_k \right)^2 \to \min$$

Система нормальных уравнений МНК

$$\partial Q / \partial \hat{\beta}_m = -2 \sum_{i=1}^{n} \left(Y_i - \sum_{k=0}^{q} x_{ik} \hat{\beta}_k \right) x_{im} = 0$$

(5) $$\sum_{i=1}^{n} x_{im} \sum_{k=0}^{q} x_{ik} \hat{\beta}_k = \sum_{i=1}^{n} x_{im} Y_i$$

$$m = 0, 1, \ldots, q$$

Свойства оценок

Обозначим $\theta = X\beta$, $\hat{\theta} = X\hat{\beta}$. Будем минимизировать величину $\varepsilon'\varepsilon = \|Y - \theta\|^2$ по отношению к $\theta \in \Omega$, где Ω – подпространство оценок $\hat{\theta}$. Если изменять значения вектора θ в пределах Ω, то квадрат длины вектора $\|Y - \theta\|^2$ достигнет минимума при значении $\theta = \hat{\theta}$, которое является проекцией вектора Y на подпространство Ω. Тогда справедливо $(Y - \hat{\theta}) \perp \hat{\theta}$ и, следовательно, $(Y - \hat{\theta}) \perp X$ (рис. 2.3). Отсюда для скалярного произведения $(Y - \hat{\theta})$ и X получаем

$$(Y - \hat{\theta}, X) = 0$$

$$X'\hat{\theta} = X'Y \quad (6)$$

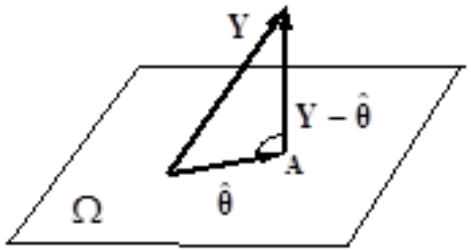


Рис. 2.3. Нахождение точки A, для которой норма минимальна

Свойства матрицы C

Обозначим **C = X′X**, , Ψ = **X′Y**, тогда СНУ запишется в виде

(7) $\mathbf{C}\boldsymbol{\beta} = \boldsymbol{\Psi}$

Свойства матрицы C:

a)так как регрессоры x1, …, xq линейно независимы, то матрица C не вырождена;

b)**C** положительно определена и ранг ее в точности равен q;

c)**C**–симметричная матрица, **C = C′**, т. е. является эрмитовой.

Поэтому СНУ (6) имеет единственное решение

(8) $\hat{\boldsymbol{\beta}} = \mathbf{C}^{-1}\boldsymbol{\Psi} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$

Оценка называется оценкой метода наименьших квадратов (МНК-оценкой). Так как в решении (8) матрица $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ неслучайная, то является линейной комбинацией наблюдений **Y**. В соответствии с теоремой Гаусса-Маркова МНК-оценка имеет наименьшую дисперсию среди всех возможных несмещенных линейных оценок.

Свойства оценок

1. **Несмещенность**

МНК-оценка $\hat{\beta}$ является случайной величиной

$$\mathbf{M}[\hat{\beta}] = \mathbf{M}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] = \mathbf{M}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \varepsilon)] =$$

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{M}[\varepsilon] = \beta.$$

2. Распределение

M[ε] = 0, D[ε] = $\sigma_\varepsilon^2\mathbf{I}_n$, или ε : $N_n(0, \sigma_\varepsilon^2\mathbf{I}_n)$, то **Y** : N_n (**X**β, $\sigma_\varepsilon^2\mathbf{I}_n$)

1) $\hat{\beta} \sim N_n(\beta, \sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1})$ 2) $(\hat{\beta}-\beta)'\mathbf{X}'\mathbf{X}(\hat{\beta}-\beta)/\sigma_\varepsilon^2 \sim \chi^2_{(q+1)}$

3) $\hat{\beta}$ **не зависит от** S_e^2 4) $SSE/\sigma_\varepsilon^2 = (n-q-1)S_e^2/\sigma_\varepsilon^2 \sim \chi^2_{(n-q-1)}$

$$S_e^2 = \frac{\mathbf{e}'\mathbf{e}}{n-q-1} = \frac{(\mathbf{Y}-\mathbf{X}\hat{\beta})'(\mathbf{Y}-\mathbf{X}\hat{\beta})}{n-q-1} = \frac{\sum_{i=1}^n e_i^2}{n-q-1} = \frac{SSE}{n-q-1}$$

где *ei* = *Yi* - $\hat{Y}$ *i* – оценка остатка ε, а SSE (Sum of Squares of Errors) – сумма квадратов (оцененных) остатков

Списки использованной литературы и источников:

- А.А.Большаков, Р.Н.Каримов «Методы обработки многомерных данных и временных рядов» Москва 2007 г.
- Электронный учебник StatSoft по анализу данных.