МАШИННЫЕ МЕТОДЫ ОБНАРУЖЕНИЯ ЗАКОНОМЕРНОСТЕЙ
(Вычислительные системы)
1981 год Выпуск 88

УДК I.I.I3.II

КЛАСС ДОПУСТИМЫХ ЭМПИРИЧЕСКИХ МОДЕЛЕЙ В ЛИНЕЙНОМ РЕГРЕССИОННОМ АНАЛИЗЕ

Э.-М.Тийт

В задачах обнаружения и моделирования закономерностей существенной проблемой является выбор признаков, входящих в модель.

В случае линейных моделей (линейного регрессионного анализа) для решения этой задачи обычно применяются разные методы пошаговой регрессии или отбора моделей (см., например, [7] или [3]), которые позволяют определить на основании выборки одну или несколько наилучших для этой выборки (эмпирически оптимальных) моделей. Но для многих прикладных задач более целесообразно определить класс "достаточно хорошо" или "допустимых" (эмпирических) моделей, среди которых самую подходящую (или самые подходящие) выбирают по содержательным соображениям. Здесь надо учитывать и эффект "инфляции" наибольшего коэффициента множественной корреляции R, указанный, например, в [8].

В настоящей заметке излагается одна возможность определения такого класса "допустимых" эмпирических моделей, описываются некоторые свойства этого класса при известных предположениях о взаимосвязях исходных признаков. На основании полученных результатов будут предложены некоторые рекомендации для практического применения регрессионного анализа в анализе данных.

§I. Определение класса допустимых эмпирических моделей

Пусть Y - функциональный (критериальный, зависимый) признак (отклик, регрессанд); $X = (X_1,\dots,X_m)$ - аргумент-признак-вектор (X_i - независимые аргументы или регрессоры), а I_p ($p \le m$) - индекс-вектор, $I_p = (i_1,\dots,i_p)$ $(1 \le i_{j-1} < i_j \le m)$, определяющий p-мерный подвектор $X(I_p) = (x_{i_1},\dots,x_{i_p})$ вектора X.

Рассматривается класс линейных регрессионных моделей (полученных методом наименьших квадратов):

$$f(X,I_p) = \sum_{i=1}^{p} a_i \cdot x_{i_1} + a_0,$$

притом качество модели характеризуется коэффициентом множественной корреляции $K(I_p)$, определяемым соотношением

$$K^2(I_p) = 1 - \frac{E(J - f(X,J_p))^2}{DJ}. \qquad (1)$$

Выборочную оценку всех величин обозначим при помощи индекса n. Теоретически наилучшей моделью является полная модель $f(X,I_m)$, где $I_m = (1,\ldots,m)$. (Предполагается, что дисперсионная матрица признак-вектора не вырождена; в противном случае существуют и другие модели, столь же хорошие как $f(X,I_m)$.) На практике может случиться, что существует подвектор $X(I_p)$ (или много таких) такой, что модель $f(X,I_p)$ при имеющейся выборке и зафиксированном уровне значимости ε не существенно хуже, чем полная модель $f(X, I_m)$. Это значит, что при заданном уровне значимости на основании имеющей выборки невозможно принять гипотезу $H_1: K(I_m) > K(I_p)$, утверждающую, что полная модель существенно лучше чем модель $f(X, I_p)$, и необходимо принять нулевую гипотезу $H_0: K(I_m) = K(I_p)$.

Все модели, которые при имеющейся выборке и заданном уровне значимости не хуже полной модели, определяют класс допустимых моделей $\mathcal{M}(\varepsilon,n)$, а соответствующие эмпирические модели, образованные на основании имеющейся выборки, - класс допустимых эмпирических моделей*) $\mathcal{M}(\varepsilon,n,x)$: $\mathcal{M}(\varepsilon,n,x) = \{f(X,I_p,n);\ H_0: K(I_p) = K(I_m)$ принимается $\}$.

Если признак Y имеет (теоретически) нормальное распределение, то для проверки H_0, и тем самым и для определения класса $\mathcal{M}(\varepsilon,n)$ (и $\mathcal{M}(\varepsilon,n,x)$) можно применять F-распределение (см. [3] и [4]). В случае более общих предположений о совместном распределении признак-вектора $(I \vdots X)$ можно применять асимптотическую методику, изложенную в статье [1].

Среди "достаточно хороших" моделей класса $\mathcal{M}(\varepsilon,n)$ целесообразно выделить "наиболее экономичные", т.е. имеющие минимальное число аргументов.

---

*) Эйткин [5] называет такие модели K^2-адекватным ε-набором.

Класс "наиболее экономичных допустимых" моделей $\mathcal{M}(\varepsilon,n,p^*)$ и число аргументов p^* для "наиболее экономичных" моделей определяются следующим соотношением:

$$\mathcal{M}(\varepsilon,n,p^*)=\{f(X,I_p);\ f(X,I_p)\in\mathcal{M}(\varepsilon,n),\ p=p^*\},$$

$$p^* = \min_{f(X,I_p)\in\mathcal{M}(\varepsilon,n)} p.$$

Классом $\mathcal{M}(\varepsilon,n,p^*)$ однозначно определяется и класс "наиболее экономичных допустимых эмпирических" моделей $\mathcal{M}(\varepsilon,n,p^*,x)$:

$$\mathcal{M}(\varepsilon,n,p^*,x) = \{f(X,I_p,n);\ f(X,I_p)\in\mathcal{M}(\varepsilon,n,p^*)\}.$$

§2. Зависимость множественного коэффициента корреляции от взаимосвязей исходных признаков

Для описания и характеризации классов $\mathcal{M}(\varepsilon,n)$ и $\mathcal{M}(\varepsilon,n,p)$ исследуем поведение коэффициента (I) при некоторых предположениях о корреляционной матрице признак-вектора $(J;X)$

$$R = \left(\begin{array}{c|c} 1 & \beta' \\ \hline \beta & R_X \end{array}\right),$$

где $\beta' = (\beta_1,\ldots,\beta_m)$, β_i – коэффициент корреляции между функциональным признаком J и регрессором X_i, а R_X есть $n\times m$-матрица корреляций аргумент-признаков. Предполагается, что R положительно определена.

Рассмотрим следующие предположения:

$$\beta_i = \beta \quad (i=1,\ldots,m), \tag{2}$$

$$R_X = R_m(\alpha) = \begin{pmatrix} 1 & \alpha & \ldots & \alpha \\ \alpha & 1 & \ldots & \alpha \\ \ldots & \ldots & \ldots & \ldots \\ \alpha & \alpha & \ldots & 1 \end{pmatrix}, \tag{3}$$

$$R_X = R_{qn}(\alpha,\delta,\gamma) = \left(\begin{array}{c|c} R_q(\alpha) & \Gamma_{qn} \\ \hline \Gamma_{nq} & R_n(\delta) \end{array}\right),\quad \Gamma_{qn} = \left.\begin{pmatrix} \overbrace{\gamma & \gamma & \ldots & \gamma}^{n} \\ \ldots & \ldots & \ldots & \ldots \\ \gamma & \gamma & \ldots & \gamma \end{pmatrix}\right\} q. \tag{4}$$

Множественный коэффициент корреляции $K(I_p)$ сравнительно просто вычисляется для следующих случаев (заметим, что рассматриваются модели, $p \le m$).

1^0. Если выполнены условия (2) и (3), то $K(I_p)$ зависит только от значений α и β и от числа регрессоров p в модели (см.[4, 5]), и имеет форму

$$K^2(I_p) = K_p^2(\alpha,\beta) = \frac{\beta\beta^2}{\alpha p - \alpha + 1} = \beta^2 H(p,\alpha) \ . \qquad (5)$$

2^0. Когда выполнены (2) и (4), то $K(I_p)$ зависит от коэффициентов корреляции β_i между J и X_i, от характеристик связей $H(k,\alpha)$ и $H(1,\delta)$ в диагональных блоках и коэффициентов корреляции γ между аргументами X_i из разных блоков:

$$K^2(I_p) = \beta^2 \frac{[H(k,\alpha)]^{-1} + [H(1,\delta)]^{-1} - 2\gamma}{[H(k,\alpha)]^{-1}\cdot[H(1,\delta)]^{-1} - \gamma} = K^2(I_p,\gamma). \qquad (6)$$

3^0. Когда (2) не выполнено, но выполнено (3), то $K(I_p)$ зависит как от α и k, так и от среднего $\bar\beta = \frac{1}{m}\sum_{i=1}^{m}\beta_i$ и от вариабельности β_i, т.е. $B = \sum_{i=1}^{m}(\beta_i - \bar\beta)^2$:

$$K^2(p) = \bar\beta^2 H(\alpha,p) + \frac{B}{1-\alpha} \ . \qquad (7)$$

Формулы (5)–(7) имеют место при следующих предположениях:

I^0. R_X положительно определена;

2^0. R положительно определена (для выполнения 2^0 необходимо выполнение I^0).

Для того чтобы выполнялось I^0 при предположении (3), необходимо и достаточно, чтобы $\alpha > -\frac{1}{p-1}$, а при выполнении (4) необходимо и достаточно, чтобы все диагональные блоки были положительно определены и еще $\gamma^2 < [H(k,\alpha)]^{-1}\cdot[H(1,\delta)]^{-1}$. Для того чтобы, выполнялось 2^0, если выполнено I^0, необходимо и достаточно, чтобы $K^2(I_p) \le 1$. Например, при предположении (2) для этого необходимо и достаточно, чтобы $\beta^2 \le \alpha + \frac{1-\alpha}{\beta}$.

§3. Доказательство формул (5)-(7)

Коэффициент множественной корреляции K^2 выражается через матрицы R_X и корреляционный вектор β формулой (см. [2, с. 237])

$$K^2 = \beta' R_X^{-1} \beta. \tag{8}$$

В случае, когда выполняется (3), обратная матрица $R_X^{-1} = (r^{ij})$ выражается следующей формулой (см. [2, с. 69]):

$$r^{ij} = \begin{cases} (\alpha m - 2\alpha+1)/((\alpha m-\alpha+1)(1-\alpha)), & \text{если } i=j, \\ -\alpha/((\alpha m-\alpha+1)(1-\alpha)), & \text{если } i\neq j \quad (i,j=1,\dots,n). \end{cases}$$

Обозначаем $\beta' R_X^{-1} = a'$, $a' = (a_1,\dots,a_m)$. Если выполнено (2), то $a_i = \beta\{[\alpha m-2\alpha+1) - (\alpha m-\alpha)]/((\alpha m-\alpha+1)(1-\alpha))\} = \beta/(\alpha m-\alpha+1)$ $(i = 1,\dots,m)$ и $K^2 = a'\beta = m\beta^2/(\alpha m-\alpha+1) = \beta^2 H(m,\alpha)$. Формула (5) доказана.

В случае, когда (3) не выполнено, получим другое выражение для a_i:

$$a_i = [\beta_i(\alpha m - 2\alpha+1) - \alpha \sum_{j\neq i} \beta_j]/((\alpha m - \alpha+1)(1-\alpha)) =$$

$$= [\beta_i(\alpha m-\alpha+1)-\alpha\Sigma\beta_j]/((\alpha m-\alpha+1)(1-\alpha)) \quad (i=1,\dots,m),$$

и соответственно

$$K^2 = a'\beta = [(\alpha m-\alpha+1)\Sigma\beta_i^2-\alpha\Sigma\beta_i\Sigma\beta_j]/((\alpha m-\alpha+1)(1-\alpha)). \tag{9}$$

Применим обозначения

$$\frac{1}{m}\sum_{i=1}^{m}\beta_i = \overline{\beta}, \qquad \sum_{i=1}^{m}(\beta_i-\beta)^2 = \sum_{i=1}^{m}\beta_i^2 - m\overline{\beta}^2 = B.$$

Подставляя их в формулу (9), получим:

$$K^2 = [(\alpha m-\alpha+1)(\Sigma\beta_i^2-m\overline{\beta}^2) + (\alpha m-\alpha+1)m\overline{\beta}^2-\alpha m^2\overline{\beta}^2]/((\alpha m-\alpha+1)(1-\alpha)) =$$

$$= \overline{\beta}^2 H(m,\alpha) + B/(1-\alpha).$$

Формула (7) доказана.

Для доказательства формулы (6) применяется выражение обратной матрицы к блочной матрице:

$$\left(\begin{array}{c|c} A & G \\ \hline G' & B \end{array}\right) = \left(\begin{array}{c|c} X & U \\ \hline U' & Y \end{array}\right), \text{ где } AX + GU' = 1,\ AU + GY = 0.$$

Отсюда получим: $X = (A - G\bar{B}^{-1}G')^{-1}$, $Y = (B - G'\bar{A}^{-1}G)$, $U = -\bar{A}^{-1}GY$. Учитывая предположение (4) и результаты вычислений, сделанных для доказательства формулы (5), получим:

$$G\bar{B}^{-1}G' = P = (p_{ij}), \quad p_{ij} = \gamma^2 H(1,\delta) \quad (i,j = 1,\ldots, q),$$

$$A - G\bar{B}^{-1}G' = S = (s_{ij}), \quad s_{ij} = \begin{cases} 1 - \gamma^2 H(1,\delta), & \text{если } i=j, \\ \alpha - \gamma^2 H(1,\delta), & \text{если } i\neq j. \end{cases}$$

Обращение матрицы S производится аналогично обращению матрицы R_X, т.е. $\bar{S}^{-1} = X = (x_{ij})$, где

$$x_{ij} = \begin{cases} [1-k\alpha+2\alpha-(k-1)\gamma^2 H(1,\delta)]/[(1+k\alpha-\alpha-k\gamma^2 H(1,\delta))(1-\alpha)], & \text{если } i=j, \\ [-\alpha+\gamma^2 H(1,\delta)]/[(1+k\alpha-\alpha-k\gamma^2 H(1,\delta))(1-\alpha)], & \text{если } i\neq j, \end{cases}$$

$i,j = 1,\ldots,l$.

И аналогично

$$y_{ij} = \begin{cases} [1-l\delta+2\delta-(l-1)\gamma^2 H(k,\alpha)]/[(1-l\delta-\delta-l\gamma^2 H(k,\alpha))(1-\delta)], & \text{если } i=j, \\ [-\delta+\gamma^2 H(k,\alpha)]/[(1+l\delta-\alpha-l\gamma^2 H(k,\alpha))(1-\delta)], & \text{если } i\neq j. \end{cases}$$

Для нахождения U вычислим $\Delta^{-1}G = C = (c_{ij})$,

$$c_{ij} = \gamma/(\alpha k - \alpha + 1) \qquad (i = 1,\ldots,k;\ j = 1,\ldots,l)$$

и, обозначая $U = (u_{ij})$, получим

$$u_{ij} = -\gamma \frac{[(1+l\delta-2\delta-(l-1)\gamma^2 H(k,\alpha)+(l-1)\gamma^2(H(k,\alpha)-H(1,\delta))]}{(\alpha k-\alpha+1)[1+l\delta-\delta-l\gamma^2 H(k,\alpha)](1-\delta)} =$$

$$= -\gamma \frac{1}{(\alpha k-\alpha+1)(\delta l-\delta+1)-lk\gamma^2}.$$

Так как $\beta' R_X \beta = \beta_1' X \beta_1 + \beta_1' U \beta_2 + \beta_1' U' \beta_2 + \beta_2' Y \beta_2$, где $\beta_1' = (\beta,\ldots,\beta)$, $\beta_2' = (\beta,\ldots,\beta)$ – соответственно k- и l-мерные постоянные векторы, имеем

$$\beta_1' X \beta_1 = \beta^2 \frac{k[1+k\alpha-2\alpha-(k-1)\gamma^2 H(1,\delta)+(k-1)(\gamma^2 H(1,\delta)-\alpha)]}{[1+k\alpha-\alpha-k\gamma^2 H(1,\delta)](1-\alpha)} =$$

$$= \beta^2 \frac{[H(1,\delta)]^{-1}}{[H(1,\delta)]^{-1}\cdot[H(k,\alpha)]^{-1}-\gamma^2}.$$

Аналогично

$$\beta_2' J \beta_2 = \beta^2 \frac{[H(k,\alpha)]^{-1}}{[H(k,\alpha)]^{-1}\cdot[H(1,\delta)]^{-1} - \gamma^2}$$

и

$$\beta_1' U \beta_2 = \beta_2' U \beta_1 = -\beta^2 \frac{\gamma}{[H(k,\alpha)]^{-1} + [H(1,\delta)]^{-1} - \gamma^2} .$$

Таким образом,

$$K^2 = \beta^2 \frac{[H(k,\alpha)]^{-1} + [H(1,\delta)]^{-1} - 2\gamma}{[H(k,\alpha)]^{-1}\cdot[H(1,\delta)]^{-1} - \gamma^2}.$$

Формула (6) доказана.

§4. Некоторые рекомендации для выбора моделей в прикладных задачах

I) Коэффициент $K^2(I_p)$, выраженный формулой (5), является монотонно возрастающим относительно β^2 и p и убывающим относительно α.

Значит, если все аргументы X_i одинаково коррелированы с функцией J (выполняется (2)), то целесообразно выбрать в модель аргументы, имеющие по возможности маленькие (или даже отрицательные) взаимные корреляции.

2) Коэффициент $K^2(I_p)$, выраженный формулой (7), монотонно зависит и от B, характеризующей вариабельность величин β_i.

Значит, если (2) не выполнено, то целесообразно выбрать в модель аргументы, имеющие разные корреляции с функцией (в том числе могут быть некоррелированные, т.е. случай $\beta_i = 0$); влияние B тем больше, чем больше взаимные корреляции аргументов.

3) При сравнительно больших α прибавление к модели $f(X,I_p)$ новых аргументов мало влияет на качество модели (в смысле $K^2(I_p)$).

4) Из формулы (6) вытекает, что блоки $R_q(\alpha)$ и $R_n(\delta)$ входят в формулу $K^2(I_p)$ симметрично, притом их размеры q и n явно в (6) не входят; значит, для выбора аргументов X_i в эти блоки следует учитывать рекомендации I и 3. В большинстве случаев положительное значение γ ухудшает модель: $K^2(I_p;\gamma_1) < K^2(I_p;\gamma_0)$, если $\gamma_0 = 0$ и $\gamma_1 > 1$. Исключением является случай, когда

$$\gamma_*^2 \approx [H(\alpha,k)]^{-1}\cdot[H(\delta,1)]^{-1},$$

тогда имеет место неравенство $K^2(I_p, \gamma_*^2) > K^2(I_p, \gamma_0)$, но такие случаи ввиду нестабильности для практики малоинтересны.

Литература

I. ПАРРИНГ А.-М. Общее асимптотическое распределение неполных множественных коэффициентов корреляции. -В кн.: Уч. зап. ТГУ, 1980, вып. 541, с. 18-26.

2. РАО С.Р. Линейные статистические методы и их применения. -М.: ИЛ, 1968.

3. СЕБЕР Дж. Линейный регрессионный анализ. -М.: ИЛ, 1980.

4. ТИЙТ Э. Выбор моделей в линейном регрессионном анализе.-В кн.: Труды ТГУ, 1981, т.46, с. 60-84.

5. ТИЙТ Э. Сравнение теоретических моделей в линейном регрессионном анализе. -Тезисы конференции "Теоретические и прикладные вопросы математики", Тарту, 1980, с. 233-235.

6. AITKIN M.A. Simultaneous inference and choice of variable subsets. -Technometrics,1974,v.16, p.221-227.

7. HOCKING R.R. The analysis and selection of variables in linear regression. - Biometrics,1976,N 32, p.1-49.

8. RENCHER A.C., FU Ceayong Pun. Inflation of R^2 in best subset regression. - Technometrics,1980,v.22, N 1,p.49-53.

Поступила в ред.-изд.отд.
28 февраля 1981 года