УДК 519.21

М. Ю. Петунина

Исследование точности модели линейной регрессии при изменении числа параметров

Изучаются точная и приближенная модели линейной регрессии для восстановления неизвестной функциональной зависимости при наличии случайных возмущений. Показано, что точность модели возрастает при уменьшении числа ее параметров.

Вивчається точна та наближена моделі лінійної регресії для відновлення невідомої функціональної залежності при наявності випадкових збурень. Показано, що точність моделі зростає при зменшенні числа її параметрів.

В приложениях математики к естествознанию и технике встречается большое число различных функциональных зависимостей, описывающих процессы, происходящие в реальной действительности. Однако получить аналитическое выражение для функции на основании априорной информации возможно далеко не во всех случаях.

В связи с этим возникает следующая задача восстановления неизвестной функциональной зависимости $y = f(x)$ на основании приближенных значений функции $f(x)$ в некоторых точках $x_1, x_2, \ldots, x_n$: определить функцию $y = f(x)$, аппроксимирующую исследуемую функциональную зависимость с заданной точностью ε. Отметим, что на практике мы почти никогда не можем наблюдать истинные значения функции $f(x)$ в точках x_i, $i = \overline{1, n}$, поскольку они всегда искажены некоторыми случайными артефактами. В простейшем случае случайные артефакты действуют аддитивно: приходим к модели, описывающей экспериментальные значения $\tilde{y}_i$ функции $f(x)$ в точках x_i

$$\tilde{y}_i = f(x_i) + \xi_i, \quad i = \overline{1, n}, \tag{1}$$

где $\{\xi_i\}$ — последовательность случайных величин, называемая шумом.

Для восстановления неизвестной функциональной зависимости $y = f(x)$ в этом случае обычно используется схема линейной регрессии [1]. Согласно схеме линейной регрессии считается, что неизвестную функцию $f(x)$ можно представить в виде линейной комбинации

$$f(x) = \sum_{j=1}^{k} c_j \varphi_j(x) \tag{2}$$

известных базисных функций $\varphi_j(x)$, $j = \overline{1, k}$ (обычно они предполагаются линейно независимыми), а коэффициенты c_j, называемые коэффициентами или параметрами линейной регрессии, считаются неизвестными и подлежат оцениванию. Для оценки этих параметров мы будем использовать классический метод наименьших квадратов (МНК) (см. [1], гл. 3). Обозначим через c_j^*, $j = \overline{1, k}$, оценки МНК параметров c_j. Тогда функцию $f(x)$ можно приближенно определить с помощью следующей формулы:

$$f^*(x) = \sum_{j=1}^{k} c_j^* \varphi_j(x). \tag{3}$$

Функцию $f^*(x)$ будем называть оценкой для $f(x)$, полученной с помощью МНК. В наиболее простом случае, когда $f(x)$ является функцией дискретного аргумента (т. е. функция $f(x)$ задана лишь в точках $x_1, x_2, \ldots, x_n$), точность оценки $f^*(x)$ можно характеризовать с помощью средней квадратической ошибки $S^2(f^*)$, определяемой по формуле

$$S^2(f^*) = \mathrm{m}\left(\sum_{i=1}^{n} (f(x_i) - f^*(x_i))^2\right). \tag{4}$$

Очевидно, что функцию $f(x)$ можно представить через базисные функции по формуле (2) многими различными способами. Основная цель этой работы — показать, что точность модели (2) возрастает при уменьшении числа параметров k схемы линейной регрессии.

1. Исследование точности модели линейной регрессии в случае независимых равноточных наблюдений. Рассмотрим вначале модель (1) в предположении, что случайные величины ξ_i являются независимыми в совокупности и имеют одинаковую дисперсию $\mathrm{D}(\xi_i) = \sigma^2$, $i = \overline{1, n}$ (независимые равноточные наблюдения по терминологии гауссовской теории ошибок). Наши дальнейшие рассуждения будут справедливы и в более общем случае, когда ξ_i представляют некоррелированные случайные величины с одинаковой дисперсией. Оказывается, что в этом случае точность модели линейной регрессии (2), определяемая средней квадратической ошибкой $S^2(f^*)$ оценки f^* функции $f(x)$, найденной по формуле (3) методом наименьших квадратов, уменьшается при увеличении числа k параметров схемы линейной регрессии (2). Точнее говоря, имеет место следующая теорема.

Теорема 1. *Пусть случайный шум $\{\xi_i\}$ в модели (1) представляет последовательность независимых в совокупности случайных величин с одинаковой дисперсией $\mathrm{D}(\xi_i) = \sigma^2$ и нулевым математическим ожиданием $\mathrm{m}(\xi_i) = 0$. Тогда средняя квадратическая ошибка оценки МНК $f^*(x_i)$ функции $f(x_i)$, $i = \overline{1, n}$, в схеме линейной регрессии*

$$f(x_i) = \sum_{j=1}^{k} c_j \varphi_j(x_i),$$

содержащей k неизвестных параметров $c_1, c_2, \ldots, c_k$, $k \leqslant n$, вычисляется по формуле $S^2(f^*) = k\sigma^2$.

Доказательство. Средняя квадратическая ошибка $S^2(f^*)$ оценки f^* имеет вид

$$S^2(f^*) = \mathrm{m}\left[\sum_{i=1}^{n}\left(\sum_{j=1}^{k} c_j^* \varphi_j(x_i) - \sum_{j=1}^{k} c_j \varphi_j(x_i)\right)^2\right] = \mathrm{m}\left[\sum_{i=1}^{n}\left(\sum_{j=1}^{k} (c_j^* - c_j)\varphi_j(x_i)\right)^2\right] =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{k}\sum_{l=1}^{k} \varphi_j(x_i)\varphi_l(x_i)\,\mathrm{m}(c_j^* - c_j)(c_l^* - c_l). \tag{5}$$

Из условий теоремы следует

$$\tilde{y}(x_i) = \sum_{j=1}^{k} c_j \varphi_j(x_i) + \xi_i, \quad i = \overline{1, n},$$

так что мы получаем следующую модель линейной регрессии:

$$\begin{pmatrix} \tilde{y}(x_1) \\ \vdots \\ \tilde{y}(x_n) \end{pmatrix} = \begin{pmatrix} \varphi_1(x_1) & \cdots & \varphi_k(x_1) \\ \vdots & \cdots & \vdots \\ \varphi_1(x_n) & \cdots & \varphi_k(x_n) \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_k \end{pmatrix} + \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix},$$

или в матричной форме $\tilde{Y} = \Phi c + \xi$.

Оценим вектор $\hat{c} = (c_1^*, c_2^*, \ldots, c_k^*)$ по методу наименьших квадратов. Имеем [2, с. 113] $\hat{c} = c + (\Phi'\Phi)^{-1}\Phi'\xi$.

Введем в рассмотрение матрицу рассеяния $V(\hat{c})$ вектора $\hat{c}$. Так как $M\hat{c} = c$, то

$$V(\hat{c}) = M(\hat{c} - c)(\hat{c} - c)' =$$

$$= \begin{pmatrix} m(c_1^* - c_1)^2 & \dots & m(c_1^* - c_1)(c_k^* - c_k) \\ \vdots & \dots & \vdots \\ m(c_k^* - c_k)(c_1^* - c_1) & \dots & m(c_k^* - c_k)^2 \end{pmatrix}. \tag{6}$$

С другой стороны [2, с. 113]

$$V(\hat{c}) = \sigma^2 (\Phi'\Phi)^{-1},$$

где

$$\Phi'\Phi = \begin{Vmatrix} \sum_{i=1}^{n} \varphi_1^2(x_i) & \dots & \sum_{i=1}^{n} \varphi_1(x_i)\varphi_k(x_i) \\ \vdots & \dots & \vdots \\ \sum_{i=1}^{n} \varphi_k(x_i)\varphi_1(x_i) & \dots & \sum_{i=1}^{n} \varphi_k^2(x_i) \end{Vmatrix}.$$

Обозначим через A определитель матрицы $\Phi'\Phi$, а через A_{jl} — алгебраическое дополнение к элементу $\sum_{i=1}^{n} \varphi_j(x_i)\varphi_l(x_i)$. Тогда

$$V(\hat{c}) = \frac{\sigma^2}{A} \begin{pmatrix} A_{11} & A_{21} & \dots & A_{k1} \\ A_{12} & A_{22} & \dots & A_{k2} \\ \cdot & \cdot & \dots & \cdot \\ A_{1k} & A_{2k} & \dots & A_{kk} \end{pmatrix}.$$

Сопоставляя полученное выражение с (6), имеем

$$m(c_j^* - c_j)(c_l^* - c_l) = \frac{\sigma^2}{A} A_{lj} = \frac{\sigma^2}{A} A_{jl}, \tag{7}$$

поскольку $(\Phi'\Phi)^{-1}$ — симметричная матрица.

Подставляя (7) в (5), получаем

$$S^2(f^*) = \sum_{l=1}^{k} \sum_{j=1}^{k} \sum_{i=1}^{n} \varphi_j(x_i)\varphi_l(x_i) \frac{\sigma^2}{A} A_{jl} = \frac{\sigma^2}{A} \sum_{l=1}^{k} \left(\sum_{j=1}^{k} \sum_{i=1}^{n} \varphi_j(x_i)\varphi_l(x_i) A_{jl} \right) =$$

$$= \frac{\sigma^2}{A} \sum_{l=1}^{k} A = k\sigma^2.$$

Теорема доказана.

Замечание. Средняя квадратическая ошибка $S^2(\tilde{Y})$ реализации процесса (1), удовлетворяющего условиям теоремы 1, вычисленная по формуле (4), равна $n\sigma^2$, причем $D(\tilde{y}_i) = \sigma^2$, $i = \overline{1, n}$. Как известно, дисперсия уменьшается при сглаживании наблюдений, т. е. при замене $\tilde{y}_i$, $i = \overline{1, n}$, на их оценки $f^*(x_i)$, найденные по методу наименьших квадратов [3, с. 103]. Кроме того, $D(f^*(x_i))$ зависит от n, и при увеличении n она уменьшается, поэтому средняя квадратическая ошибка $S^2(f^*)$, которая равна $\sum_{i=1}^{n} D(f^*(x_i))$, может не зависеть от n (и как показывает теорема 1, она действительно не зависит от n), поскольку увеличение числа слагаемых компенсируется уменьшением их абсолютных значений. Из теоремы 1 также вы-

текает, что точность сглаженной функции $f^*(x)$ в n/k, $k \leqslant n$, раз превышает точность реализации $\tilde{Y}$.

Непосредственно из теоремы 1 вытекает следующее утверждение.

Следствие 1. *В условиях теоремы 1 с увеличением числа параметров линейной регрессионной модели* (2) *ее средняя квадратическая ошибка* $S^2(f^*)$ *возрастает, поэтому точность этой модели уменьшается.*

2. Исследование точности приближенной линейной регрессионной модели. В предыдущем пункте изучалась схема линейной регрессии, в которой неизвестную функцию $f(x)$, подлежащую оцениванию, можно было представить в виде линейной комбинации базисных функций $\varphi_j(x)$. Эта ситуация довольно далека от реальной действительности, поскольку неизвестная функция $f(x)$ обычно разлагается в ряд по базисным функциям $\varphi_j(x)$:

$$f(x) = \sum_{j=1}^{\infty} c_j \varphi_j(x)$$

и ее нельзя представить в виде их конечной линейной комбинации. В связи с этим рассмотрим следующую схему линейной регрессии, которую мы будем называть приближенной линейной регрессионной моделью: линейное многообразие

$$\mathcal{L} = \Big\{\varphi(x) : \varphi(x) = \sum_{j=1}^{k} c_j \varphi_j(x), \quad c_j \in R^1, \quad j = \overline{1, k}\Big\},$$

порожденное базисными функциями $\varphi_1(x), \ldots, \varphi_k(x)$, не содержит неизвестной оцениваемой функции $f(x)$. Таким образом, приближенная модель линейной регрессии, в отличие от точной, рассмотренной в п. 1, характеризуется тем, что неизвестная функция $f(x)$ отстоит от линейного многообразия $\mathcal{L}$ на положительном расстоянии δ в некоторой метрике ρ:

$$\delta = \rho(f, \mathcal{L}) = \inf_{\varphi \in \mathcal{L}} \rho(f, \varphi) > 0.$$

В качестве метрики ρ будем использовать, как обычно, гильбертову метрику пространства L_2: для двух функций $f(x_i)$ и $\varphi(x_i)$, $i = \overline{1, n}$, дискретного аргумента x гильбертова метрика определяется формулой

$$\rho(f, \varphi) = \Big(\sum_{i=1}^{n} (\varphi(x_i) - f(x_i))^2\Big)^{1/2}.$$

Квадрат расстояния $\delta = \rho(f, \mathcal{L})$ в гильбертовой метрике будем называть аппроксимационной ошибкой приближенной модели линейной регрессии.

Точность приближенной модели линейной регрессии также будем характеризовать средней квадратической ошибкой $S^2(f^*)$, однако в приближенной модели, в отличие от точной, средняя квадратическая ошибка $S^2(f^*)$ состоит из двух компонент, одну из которых представляет аппроксимационная ошибка δ^2 (детерминированная часть средней квадратической ошибки), а вторая компонента обусловлена ошибками, связанными с наличием возмущающего воздействия (шума) $\xi(x)$. Точнее говоря, имеет место следующая теорема.

Теорема 2. *Пусть случайный шум* $\{\xi_i\}$ *приближенной модели с аппроксимационной ошибкой* δ^2

$$\tilde{y}_i = f(x_i) + \xi_i = \sum_{j=1}^{k} c_j \varphi_j(x_i) + r(x_i) + \xi_i, \quad i = \overline{1, n}, \quad \|R\|^2 = \sum_{i=1}^{n} |r(x_i)|^2 = \delta^2,$$

$$R = \begin{pmatrix} r(x_1) \\ \vdots \\ r(x_n) \end{pmatrix} \tag{8}$$

представляет последовательность независимых в совокупности случайных величин с одинаковой дисперсией $D(\xi_i) = \sigma^2$ *и нулевым математическим ожи-*

данием $\mathrm{m}(\xi_i) = 0$. *Тогда средняя квадратическая ошибка* $S^2(f^*)$ *оценки* $f^*(x_i)$ *функции* $f(x_i)$, $i = \overline{1, n}$, *в приближенной схеме линейной регрессии* (8), *содержащей* k *неизвестных параметров* $c_1, c_2, \ldots, c_k$, $k \leqslant n$, *вычисляется по формуле* $S^2(f^*) = k\sigma^2 + \delta^2$.

Доказательство. Легко видеть, что приближенная модель линейной регрессии может быть записана в эквивалентной матричной форме следующим образом: $\tilde{Y} = \Phi c + R + \xi$. При этом $R = Z\gamma$, где

$$Z = \begin{Bmatrix} z_1(x_1) & \ldots & z_{n-k}(x_1) \\ \cdot & \ldots & \cdot \\ \cdot & \ldots & \cdot \\ \cdot & \ldots & \cdot \\ z_1(x_n) & \ldots & z_{n-k}(x_n) \end{Bmatrix}, \qquad \gamma = \begin{Bmatrix} \gamma_1 \\ \cdot \\ \cdot \\ \cdot \\ \gamma_{n-k} \end{Bmatrix}.$$

Действительно, столбцы матрицы Z представляют базисные векторы подпространства, ортогонального к линейному многообразию $\mathcal{L}$, а величина $\| R \|^2$ достигает своего минимума, когда $R \perp \mathcal{L}$, поэтому вектор R принадлежит подпространству $\mathcal{L}^\circ$, ортогональному к $\mathcal{L}$. Следовательно, вектор R можно представить в виде линейной комбинации базисных векторов подпространства $\mathcal{L}^\circ$, ортогонального к $\mathcal{L}$: $R = Z\gamma$.

По условию теоремы $M\xi = 0$, поэтому $M\tilde{Y} = \Phi c + Z\gamma = f$, где

$$f = \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix}.$$

Положим $\hat{f} = \Phi\hat{c}$. Средняя квадратическая ошибка приближенной модели (8) имеет вид

$$S^2(f^*) = \mathrm{m}(f - \hat{f})'(f - \hat{f}) = \mathrm{m}(\Phi c + Z\gamma - \Phi\hat{c})'(\Phi c + Z\gamma - \Phi\hat{c}) =$$

$$= \mathrm{m}[(\Phi c - \Phi\hat{c})'(\Phi c - \Phi\hat{c}) + (Z\gamma)'(\Phi c - \Phi\hat{c}) + (\Phi c - \Phi\hat{c})' Z\gamma +$$

$$+ (Z\gamma)' Z\gamma] = \mathrm{m}(\Phi c - \Phi\hat{c})'(\Phi c - \Phi\hat{c}) + \delta^2,$$

поскольку вектор R в силу условий (8) ортогонален подпространству $\mathcal{L}$, а вектор $\Phi c - \Phi\hat{c} \in \mathcal{L}$.

Так как

$$(\Phi c - \Phi\hat{c})'(\Phi c - \Phi\hat{c}) = \sum_{i=1}^{n} \Big(\sum_{j=1}^{k} (c_j - c_j^*) \varphi_j(x_i) \Big)^2,$$

то

$$\mathrm{m}(\Phi c - \Phi\hat{c})'(\Phi c - \Phi\hat{c}) = \sum_{i=1}^{n} \sum_{j=1}^{k} \sum_{l=1}^{k} \varphi_j(x_i)\, \varphi_l(x_i)\, \mathrm{m}(c_j^* - c_j)(c_l^* - c_l).$$

Найдем вектор $\hat{c}$ для этого случая:

$$\hat{c} = (\Phi'\Phi)^{-1}\Phi'\tilde{Y} = (\Phi'\Phi)^{-1}\Phi'(\Phi c + Z\gamma + \xi) = c + (\Phi'\Phi)^{-1}\Phi' Z\gamma +$$

$$+ (\Phi'\Phi)^{-1}\Phi\xi = c + (\Phi'\Phi)^{-1}\Phi\xi,$$

поскольку $\Phi' Z = 0$.

Мы видим, что вектор $\hat{c}$ имеет такой же вид, как и в теореме 1. Следовательно, его матрица рассеяния равна $\sigma^2(\Phi'\Phi)^{-1}$, поэтому проводя рассуждения, аналогичные проведенным при доказательстве теоремы 1, получаем

$$\mathrm{m}(\Phi c - \Phi\hat{c})'(\Phi c - \Phi\hat{c}) = k\sigma^2,$$

так что $S^2(f^*) = k\sigma^2 + \delta^2$. Теорема доказана.

Из теоремы 2 немедленно вытекает такое утверждение.

Следствие 2. *Точность приближенной линейной регрессионной модели* (8), *удовлетворяющей условиям теоремы 2, уменьшается, когда аппроксимационная ошибка модели остается постоянной, а число ее параметров увеличивается.*

1. *Себер Дж.* Линейный регрессионный анализ.— М. : Мир, 1980.— 456 с.
2. *Кендалл М. Дж., Стьюарт А.* Статистические выводы и связи.— М. : Наука, 1973.— 899 с.
3. *Волков Е. А.* Численные методы.— М. : Наука, 1982.— 254 с.

Ин-т математики АН УССР, Киев

Получено 17.12.87