УДК 519.2

ОКРУГЛЕНИЕ ЧИСЛЕННЫХ ПАРАМЕТРОВ РЕГРЕССИОННОЙ МОДЕЛИ

к.т.н. В.Ю. Дубницкий, к.т.н. Н.А. Цейтлин*
(представил д.ф.-м.н. С.В. Смеляков)

Представлен способ статистически обоснованного округления регрессионных коэффициентов многофакторных моделей.

Большинство задач экспериментального исследования завершается построением регрессионной модели (РМ) исследуемого объекта. Обычно численные значения оценок $\mathbf{b_i}$ коэффициентов $\boldsymbol{\beta_i}$ и оценки $\mathbf{S^2}$ остаточной дисперсии $\boldsymbol{\sigma^2}$ записываются с чрезмерно большим количеством (от 4 до 8) цифр, что приводит к ошибочному представлению о большой точности оценок и увеличивает объем вычислительных работ с незначащими цифрами. Необходимость создания статистически обоснованных правил округления численных значений оценок статистических параметров отмечена в [1]. Необоснованное округление численных значений $\mathbf{b_i}$ приводит к тому, что остаточная ошибка $\boldsymbol{\varepsilon \sim N(0,\sigma^2)}$ для РМ с неокругленными коэффициентами превращается в существенно худшую ошибку $\boldsymbol{\varepsilon_0 \sim N(y_0,\sigma_0{}^2)}$ для РМ с округленными значениями, появляется значимая систематическая ошибка $\boldsymbol{\nu_0 \neq 0}$ и $\boldsymbol{\sigma_0{}^2 > \sigma^2}$. В [2] получено правило округления коэффициентов РМ для одной переменной. Авторами получено решение этой задачи для случая нескольких переменных.

Для решения требуется определение зависимости количества значащих цифр $\mathbf{u}$, на которое надо увеличить оценки коэффициентов РМ по сравнению с числом значащих цифр, которое предписывается правилом округления одной оценки, от статистических характеристик РМ. Приведем обоснование правила округления одной оценки $\hat{\theta}$ параметра θ.

Допустим, что после округления оценки $\hat{\theta}$ параметра θ получится оценка $\hat{\theta}_0$ другого параметра $\theta_0 \neq \theta$. Это позволяет сформулировать нулевую гипотезу $\mathbf{H_0}$ о том, что неокругленная ($\hat{\theta} \to \theta$) и округленная ($\hat{\theta}_0 \to \theta$) оценки однородны, т.е. $\mathbf{H_0}$: $\theta - \theta_0 = \lambda = 0$ против альтернативной $\mathbf{H_1}$: $\lambda \neq 0$. Проверочную статистику запишем в виде $\mathbf{z = \hat{\lambda} / S_{\hat{\lambda}}}$, где $\hat{\lambda}$ - оценка разности $\hat{\lambda} = \hat{\theta} - \hat{\theta}_0$; $\mathbf{S_{\hat{\lambda}}}$ - среднеквадратичное отклонение (СКО) ошибки разности.

Для построения правила округления оценок с большим запасом точности введем следующие допущения. Положим, что $S = S_{\hat{\theta}}$, где $S_{\hat{\theta}}$ - СКО оценки $\hat{\theta}$; величина $z = \hat{\lambda} / S_{\hat{\lambda}}$ подчинена нормальному закону $z \sim N(0,1)$. Принимая ответственность за вывод "предельно большую" и учитывая, что "предпочтительной" является гипотеза H_0, согласно рекомендациям, приведенным в [3], зададим очень большое критическое значение уровня значимости $\alpha_{k0} = 0,6$ (этому числу соответствует квантиль $z_{\alpha/2}$ функции нормированного нормального распределения $N(0,1)$, соответствующий вероятности $p = (1-\alpha)/2 = (1-0,6)/2 = 0,2$).

При $P=0,2$ соответствующий квантиль будет равен 0,5; т.е. $z_{0,6/2} = 0,5$. Тогда условие, при котором гипотеза H_0 не отклоняется, имеет вид

$$z_{0,3} = 0,5 > \hat{z} = |\hat{\theta} - \hat{\theta}_0| / S_{\hat{\theta}}. \qquad (1)$$

Этому условию удовлетворяет следующее правило [2]: для округления абсолютного численного значения оценки $\hat{\theta}$ необходимо в числе $\hat{\theta}$ отбросить цифры младшего разряда, начиная с цифры, соответствующей второй цифре наибольшего разряда среднеквадратичного отклонения оценки $S_{\hat{\theta}}$. Последнюю из оставшейся в округленной оценке $\hat{\theta}_0$ цифру увеличивают на единицу, если первая из отбрасываемых цифр больше или равна 5.

Пример 1. Пусть СКО $S_{\hat{\theta}} = 0,024$. Тогда оценку $\hat{\theta} = 1,068553$ округляют до $\hat{\theta}_0 = 1,06$, а оценку $\hat{\theta} = 1,995047$ - до $\hat{\theta}_0 = 2,00$.

Теперь приведем обоснование правила округления оценок коэффициентов РМ в многомерном случае.

Пусть по известной матрице плана $\underline{X}$ размерности $(N \times k)$ и вектору откликов $\underline{Y} = (Y_1, Y_2, ..., Y_N)$ методом наименьших квадратов получены неокругленные численные оценки b_i (в большинстве программ ЭВМ используется от 6-ти до 9-ти цифр) коэффициентов $\beta_i \left(i = \overline{1,m}\right)$ РМ $\hat{Y} = \hat{f}(x,b)$; СКО ошибок этих коэффициентов S_{bi}, оценка СКО остаточной ошибки σ равна s.

Необходимо округлить численные оценки b_i коэффициентов РМ.

Допустим, что после округления оценок b_i получены оценки b_{i0} других параметров β_{i0}, приводящих к смещению центра распределения ошибки ε модели регрессии $Y = f(x,b) + \varepsilon$ на величину $\nu_0 \neq 0$ и дисперсии σ^2 на величину $\sigma_a^2 > 0$. Это позволяет сформулировать одну пару гипотез относительно математического ожидания ошибки:

$$H_{10}: \nu_0 = 0;\ H_{11}: \nu_0 \neq 0. \qquad (2)$$

Для формулирования второй пары гипотез относительно дисперсии величины ε будем считать, что приращение σ_a^2 дисперсии σ^2 пренебрежимо

мало по сравнению с дисперсией σ^2, если величина σ_a^2 не превышает дисперсии σ_S^2 СКО **S**. Приходим к гипотезам:

$$\mathbf{H_{20}: \sigma_a^2 \le \sigma_S^2;}\ .\ \mathbf{H_{21}: \sigma_a^2 > \sigma_S^2}\ . \qquad (3)$$

Объединяя гипотезы (2) и (3), приходим к одной нулевой гипотезе $\mathbf{H_0}$, являющейся конъюнкцией $\mathbf{H_{10}}$ и $\mathbf{H_{20}}$, против альтернативы $\mathbf{H_1}$, являющейся дизъюнкцией $\mathbf{H_{11}}$ и $\mathbf{H_{21}}$:

$$\mathbf{H_0: H_{10} \cap H_{20};\ H_1: H_{11} \cap H_{21}}\ . \qquad (4)$$

Как и прежде, для построения правила округления оценок с большим запасом точности возьмем в качестве статистической характеристики гипотезы $\mathbf{H_{10}}$ отношение величины $\bar{\mathbf{e}}_0$ - среднего значения остатков после округления коэффициентов РМ к величине $\mathbf{S_{\bar{e}}}$ – СКО ошибки среднего $\bar{\mathbf{e}}$ остатков РМ: $\mathbf{\bar{e}/S_{\bar{e}}=z;\ z\sim N(0,1).}$

Статистика $\mathbf{S^2}$ после округления коэффициентов примет значение $\mathbf{S_0^2}$ с $\mathbf{f_0}$ степенями свободы. Оценку $\mathbf{S_a^2}$ дисперсии σ_a^2 находим по формуле $\mathbf{S_a^2=S_0^2 - S^2}$. Оценку $\mathbf{S_s^2}$ дисперсии σ_s^2 находим по формуле $\mathbf{S_s^2=S^2/\ (2f)}$ с числом степеней свободы **f**. Ввиду независимости дисперсий $\mathbf{S_a^2}$ и $\mathbf{S_s^2}$, в качестве статистической характеристики $\mathbf{H_{20}}$ можно выбрать фишеровское отношение с $\mathbf{f_a}$ и **f** степенями свободы

$$\mathbf{\bar{F} = S_a^2 / S_s^2}\ . \qquad (5)$$

Зададим очень большое значение критического уровня значимости $\mathbf{\alpha_{k0}=0,75}$. Всего имеется две альтернативные гипотезы (4). Можно показать [3], что условие, при котором гипотеза $\mathbf{H_0}$ не отклоняется, имеет вид

$$\mathbf{\hat{\alpha} = 1 - \left(1 - \min\left\{\hat{\alpha}_z, \hat{\alpha}_F\right\}\right)^2 \ge \alpha_{k0}}\ , \qquad (6)$$

где $\mathbf{\hat{\alpha}_z, \hat{\alpha}_F}$ - оценки уровней значимости статистик $\mathbf{\hat{z}}$ и $\mathbf{\hat{F}}$.

Зададим $\mathbf{\hat{\alpha}_z = \hat{\alpha}_F = \tilde{\alpha}}$ и решим неравенство (6) относительно $\mathbf{\hat{\alpha}}$. Тогда

$$\mathbf{\tilde{\alpha} \ge 1 - \left(1 - \alpha_{k0}\right)^{1/2} = \tilde{\alpha}_{k0}}\ . \qquad (7)$$

Подставив сюда $\mathbf{\alpha_{k0} = 0,75}$, получим критическое значение уровня значимости для проверки каждой из двух гипотез: $\mathbf{\hat{\alpha}_{k0} = 0,5}$. Видно, что число $\mathbf{\alpha_{k0} = 0,75}$ было выбрано так, чтобы критические значения критериев $\mathbf{z_k}$ и $\mathbf{F_k}$ можно было задать достаточно малыми: $\mathbf{z_k \approx 0,7}$ и $\mathbf{F_k \approx 1}$ (для всех **f** и $\mathbf{f_0}$). Это позволяет записать условия, при которых гипотеза $\mathbf{H_0}$ не отклоняется:

$$\mathbf{\hat{z} = \left|\overline{e_0}\right| / S_{\bar{e}} = \left\{\left|\overline{e_0}\right| N^{0,5} / S\right\} < Z_k = 0,7}\ ; \qquad (8)$$

$$\mathbf{\hat{F} = S_a^2 / S_s^2 = 2f\left(S_0^2 / S^2 - 1\right) < F_k = 1}\ ; \qquad (9)$$

Теперь опишем процедуру округления коэффициентов РМ.

1. Установить число значащих цифр в коэффициентах РМ по сформулированному выше правилу округления одной единственной оценки (следует использовать значения коэффициентов $\mathbf{b_i}$ и СКО их ошибок $\mathbf{S_{bi}}$).

2. Определить такое целое число **U**, на которое следует дополнительно увеличить число значащих цифр в записи чисел всех коэффициентов РМ. Начать с нулевого приближения **(U=0)**. Найти остатки $\mathbf{e_{iu} = Y_i - f\left(\underline{B}_{ou}, \underline{x}_i\right)}$, где $\underline{\mathbf{B}}_{\mathbf{ou}}$ – вектор округленных значений коэффициентов РМ; $\underline{\mathbf{Y}}_{\mathbf{i}}$ – координаты вектора $\underline{\mathbf{Y}}$; $\mathbf{i = \overline{1,n}}$.

3. Вычислить значения составляющих остаточной ошибки: среднее $\bar{\mathbf{e}}_{\mathbf{ou}} = \sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{e_{iu}} \big/ \mathbf{N}$ и СКО $\mathbf{S_{ou}} = \left[\sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{e_{iu}^2} \big/ \left(\mathbf{N - m}\right)\right]^{\mathbf{0,5}}$.

4. Проверить справедливость (8) и (9). Если хотя бы одно из этих условий не удовлетворено, то увеличить число значащих цифр в округляемых коэффициентах РМ еще на одно (заменить **U** на **U+1**) и возвратиться к проверке (8) и (9). Если (8) и (9) удовлетворены, то выбор числа **U** завершен.

Пример 2. Округлить численные оценки параметров РМ $\hat{\mathbf{Y}} = \mathbf{-0{,}963837 + 1{,}51714 * X_1 - 0{,}606018 * X_2}$ с вектором СКО ошибок коэффициентов $\underline{\mathbf{S}}_{\mathbf{b}} = \left(\mathbf{1{,}51260; 0{,}0585829; 0{,}0642270}\right)^{\mathbf{T}}$ и СКО остаточной ошибки **S=1,15762**. Известны $\underline{\mathbf{X}}$ и $\underline{\mathbf{Y}}$: $\underline{\mathbf{X}} = \begin{pmatrix} 31 & 40 \\ 76 & 73 \\ 64 & 63 \\ 70 & 52 \\ 28 & 20 \end{pmatrix}$; $\underline{\mathbf{Y}} = \begin{pmatrix} 21 \\ 70 \\ 59 \\ 73 \\ 30 \end{pmatrix}$.

Решение. Выразим (8) и (9) относительно $\bar{\mathbf{e}}_{\mathbf{0}}$ и $\mathbf{S_0}$ (при **N=5** и **f=N-m=2**):

$$\left|\bar{\mathbf{e}}_{\mathbf{0}}\right| < \mathbf{0{,}7S} \big/ \mathbf{N^{0,5}} = \mathbf{0{,}362}\ ;\ \mathbf{S_0 < S\left(1 + 1/\left(2f\right)\right)^{0,5} = 1{,}29}\ . \qquad (10)$$

Установим в нулевом приближении **(U=0)** число значащих цифр в коэффициентах РМ по правилу округления одной оценки. С помощью полученной РМ $\mathbf{Y=-1+1{,}52X_1-0{,}61X_2}$ найдем остатки $\mathbf{e_{io} = \left(-0{,}72; 0{,}01; 1{,}15; -0{,}68; 0{,}64\right)^T}$, а по ним – среднее $\bar{\mathbf{e}}_{\mathbf{00}} \mathbf{= 0{,}08}$ и СКО $\mathbf{S_{00} = 1{,}16}$. Полученные значения $\bar{\mathbf{e}}_{\mathbf{00}}$ и $\mathbf{S_{00}}$ удовлетворяют условиям (10).

Ответ: вектор округленных значений коэффициентов регрессионной модели $\underline{\mathbf{B}}_{\mathbf{0}} = \left(\mathbf{-1\,1{,}52 - 0{,}61}\right)^{\mathbf{T}}$; СКО остаточной ошибки $\mathbf{S_0 = 1{,}2}$.

ЛИТЕРАТУРА

1. Шеффе Г. Дисперсионный анализ. – М.: Наука, 1980. – 517 с.
2. Урбах В.Ю. Биометрические методы. – М.: Наука, 1964. – 380 с.
3. Цейтлин Н.А. Применение методов математической теории эксперимента в содовой промышленности. – М.: НИИТЭХИМ. 1984. – 36 с.