УДК 519.23

ЭФФЕКТИВНЫЙ МЕТОД РАНЖИРОВАНИЯ НЕЗАВИСИМЫХ ПЕРЕМЕННЫХ И ОТБРАСЫВАНИЯ НЕСУЩЕСТВЕННЫХ ПАРАМЕТРОВ ПРИ МНОГОФАКТТОРНОМ СТАТИСТИЧЕСКОМ АНАЛИЗЕ[1]

© Поляков Б.Н.

e-mail: *bpoliakov@hotmail.com*

Abstract. The reliable criteria of ranging of independent variables and rejection of insignificant parameters at the multifactorial statistical analysis substantiations are resulted and are offered, which efficiency is illustrated by a concrete example and proves to be true more than 30-years practice of successful carrying out of statistical researches in mechanical engineering, metallurgy and medicine.

Введение

В работе [1] отмечалось, что для нахождения криволинейного уравнения множественной регрессии методом Брандона, независимые переменные необходимо ранжировать, т.е. располагать последние в порядке уменьшения силы их влияния на зависимую переменную. Ранжировать независимые переменные можно на основе линейного регрессионного анализа, как первого этапа многофакторного анализа, несколькими способами: по коэффициенту полной корреляции d_{1i}, по коэффициенту частной корреляции r_{1i} или по стандартизованному коэффициенту регрессии β_i.

Коэффициент полной корреляции d_{1i} характеризует тесноту связи между зависимой переменной x_1 и независимой x_i вне зависимости от того, чем обусловлена эта связь, действительным влиянием x_i, либо влиянием других независимых переменных, корреляционно связанных с x_i и, вследствие этого, искажающих силу влияния рассматриваемой независимой переменной на зависимую. Особенно это отмечается в том случае, когда независимые переменные сильно коррелируют между собой.

Стандартизованный коэффициент регрессии β_i является количественной характеристикой силы влияния независимой переменной, выраженной в единицах среднеквадратического отклонения зависимой переменной при устранении другой линейной связи с остальными независимыми переменными. Но в анализе связи двух переменных обычно принято пользоваться не абсолютными оценками, а относительными, т.е. более общими характеристиками.

Коэффициент частной корреляции r_{1i} является относительной характеристикой силы влияния независимой переменной на зависимую при постоянстве других, участвующих в анализе, т.е. выражает влияние, очищенное от действия других независимых факторов. Из определения коэффициента частной корреляции ясно, что последний является наилучшей оценкой силы связи между независимой и зависимыми переменными на основе линейного приближения к эмпирическим данным. И поэтому принято в последовательном многофакторном анализе [1] независимые переменные располагать по мере убывания $|r_{1i}|$.

[1]Разработан совместно с канд.техн.наук Ю.Д. Макаровым и инж.- математиком Ф.М.Карлинской

1. КРИТЕРИИ РАНЖИРОВАНИЯ НЕЗАВИСИМЫХ ПЕРЕМЕННЫХ И ОТБРАСЫВАНИЯ НЕСУЩЕСТВЕННЫХ ПАРАМЕТРОВ

Рассмотрим теперь вопрос отбрасывания несущественных параметров. Обычно исследователь стремится зафиксировать как можно больше независимых переменных, которые, по его мнению, каким-то образом влияют на изучаемое явление. При этом многие «независимые» факторы могут быть на самом деле тесно взаимосвязаны друг с другом. Большое количество независимых переменных иногда искажает физический смысл определяемого уравнения, к тому же коэффициенты регрессии, вычисленные для сильно коррелированных переменных, малонадежны и будут иметь широкие доверительные интервалы, поэтому часть независимых переменных на основе какого-либо критерия необходимо исключить из рассмотрения. Таким образом, возникает вопрос о критерии отбрасывания несущественных параметров.

О существенности влияния независимой переменной на зависимую переменную можно судить уже по величине доверительного интервала коэффициента регрессии. Если доверительный интервал проходит через нуль, то вопрос о существенности влияния соответствующей независимой переменной ставится под сомнение. Поэтому выполнение неравенства (1) будет являться естественным критерием существенности независимого фактора

$$t_{iI} = \left| \frac{a_i}{S_{a_i}} \right| > t_\alpha, \tag{1}$$

где a_i – коэффициент регрессии, S_{a_i} – его среднеквадратическое отклонение, t_α – квантиль нормированного нормального распределения, соответствующий вероятности $(1-\alpha)$. Данный критерий (неравенство (1)) будем называть первым, на что указывает римская цифра I у индекса в обозначении критерия $t_{\alpha I}$. Но очень часто, по многим причинам: неправильно выбраны независимые параметры, недостаточны величина выборки и точность экспериментальных данных, и вследствие этого, небольшая точность результатов анализов и т.д., – почти для всех коэффициентов регрессии несправедливо неравенство (1) или доверительные интервалы этих коэффициентов проходят через нуль. И одновременное отбрасывание несущественных независимых переменных по критерию I может резко уменьшить коэффициент множественной корреляции, увеличить стандартную ошибку оценки и вообще привести к неправильному объяснению изучаемого процесса.

Число же всевозможных вариантов отбрасывания независимых переменных на основе критерия I растет по закону $C_k^1 + C_k^2 + ... + C_k^{k-1} + C_k^k$ где k – количество независимых переменных, которые нужно было бы отбросить по критерию I.

В. Визорке и др. [2] предложили метод одновременного отбрасывания нескольких независимых переменных на основе собственного опыта, полученного при обработке большого количества экспериментальных данных. Суть этого метода в следующем.

Рассмотрим величину $t_i = a_i/S_{a_i}$, эквивалентной формулой которой будет являться следующая:

$$t_i = \frac{\beta_i \sqrt{1 - R_i^2}}{\sqrt{1 - R^2}} \sqrt{n - m},$$

где β_i – стандартизованный коэффициент регрессии, R – коэффициент множественной корреляции, R_i – коэффициент множественной корреляции i – той независимой переменной с остальными независимыми переменными, n – количество значений зависимой переменной, m – число рассматриваемых параметров, включая зависимый. Если x_i коррелирует только с одной независимой переменной, и если последняя является несущественной по критерию I, то после ее отбрасывания t_i возрастает в $1/\sqrt{1 - R_i^2}$ раз, что не учитывает рассмотренный выше критерий.

Поэтому необходимо относиться очень осторожно к независимым переменным, которые тесно взаимосвязаны с другими независимыми переменными, и в первую очередь, рекомедуется отбросить те независимые переменные, несущественные по критерию I , коррелирующие слабо с другими, так как данный критерий не учитывает взаимной корреляции между независимыми переменными. Но независимая переменная может коррелировать с несколькими независимыми переменными и t_i может возрасти [2] приблизительно в $1 + R_i\sqrt{2}/\sqrt{1 - R_i^2}$ раз после отбрасывания несущественных параметров и

$$t_{iII} = |t_{iI}| \frac{1 + R_i\sqrt{2}}{\sqrt{1 - R_i^2}} > t_\alpha, \tag{2}$$

Экспериментальная проверка показала, что критерий II отбрасывания нескольких независимых параметров очень слаб, что будет проиллюстрировано в приведённом ниже примере. Для практического применения этого критерия можно его усилить. В алгоритме многофакторного статистического анализа [1] заложен следующий критерий:

$$t_{iIII} = |t_{iI}| \frac{1}{\sqrt{1 - R_i^2}} > t_\alpha, \tag{3}$$

Критерий отбрасывания II отличается от критерия III сомножителем $1 + R_i\sqrt{2}$ в числителе правой части выражения (2).

На Рис. 1. и Рис. 2. показаны зависимости $1+R_i\sqrt{2}/\sqrt{1 - R_i^2}$ и $1/\sqrt{(1 - R_i^2)}$ от R_i. Из рассмотрения этих графиков и неравенств (2) и (3) можно сделать вывод о том, что в первую очередь будут отброшены те независимые переменные, для которых мало значение R_i, т.е. слабо связанные с другими независимыми переменными.

После работы критерия II могут остаться ещё несколько независимых переменных существенных по критерию III, но не существенных по критерию I. В этом случае предлагается на втором этапе исследования несущественные переменные отбрасывать по следующему критерию:

$$t_{iIV} = |t_{iI}| \frac{1 + |r_{1i}|\sqrt{2}}{\sqrt{1 - r_{1i}^2}} > t_\alpha, \tag{4}$$

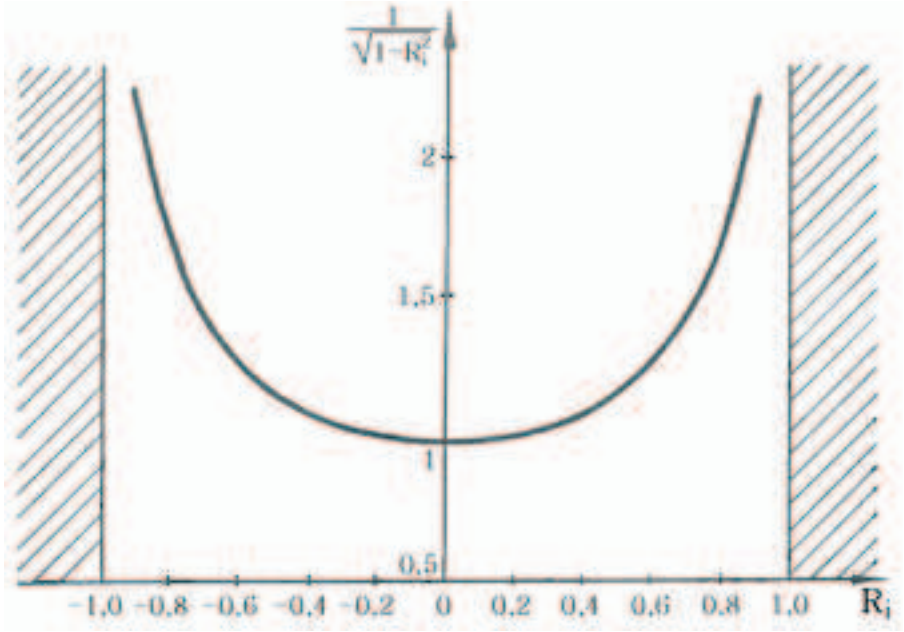

Рис. 1. Зависимость $\frac{1}{\sqrt{1-R_i^2}}$ от R_i

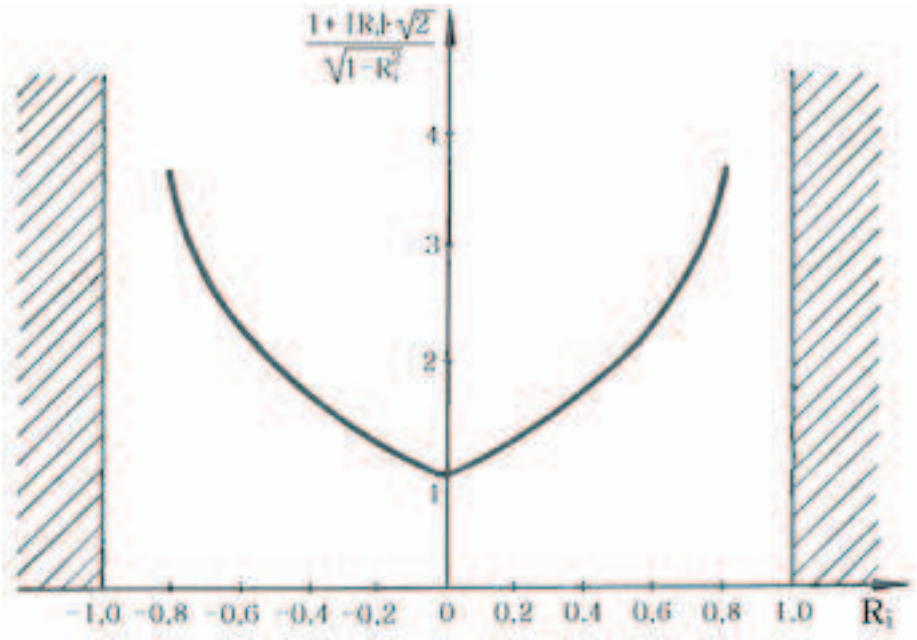

Рис. 2. Зависимость $\frac{1+|R_i|\sqrt{2}}{\sqrt{1-R_i^2}}$ от R_i

Сопоставляя критерии отбрасывания независимых переменных II и IV, можно заметить, что они имеют одинаковую структуру. Поэтому график функции

$1 - R_i\sqrt{2}/\sqrt{1-R_i^2}$ полностью совпадает с графиком функции $1 + |r_{1i}|\sqrt{2}/\sqrt{1-r_{1i}^2}$. Из Рис. 2 и неравенства (4) следует, что в первую очередь будут отбрасываться те несущественные независимые переменные, которые слабо влияют на зависимую переменную. Практическое применение критерия IV дало положительный результат. Критерий IV сильнее критерия III, поэтому он применяется уже на втором этапе отбрасывания независимых переменных.

2. Пример

Действия приведенных выше критериев проиллюстрированы примером. Опытные данные для примера взяты из источника [3], в котором обобщены технологические режимы прокатки на блюмингах ряда отечественных металлургических заводов и дополнены данными по схемам и режимам обжатий, применяющимися на блюминге 1300 меткомбината «Криворожсталь».

Пример. На основе опытных данных определяется регрессионная зависимость количества пропусков (N) от среднего обжатия в пропуске за цикл прокатки ($\Delta_{\text{ср}}$, мм), площади поперечного сечения слитка ($F_{\text{сл}}$, мм^2), площади поперечного сечения конечной заготовки ($F_{\text{пр}}$, мм^2), числа кантовок (K), массы слитка ($Q_{\text{сл}}$, т), числа пропусков до первой кантовки (n_{K1}), числа пропусков между первой и второй кантовкой (n_{K2}), числа пропусков между второй и третьей кантовкой (n_{K3}), числа пропусков между третьей и четвертой кантовкой (n_{K4}).

Таблица 1. Первое приближение

Параметр	a_i	$a_i^{(1)}$	$a_i^{(2)}$	r_{1i}	t_{iI}	t_{iII}	t_{iIII}	t_{iIV}	$\sqrt{1-R_i^2}$
N	13,81	13,99	13,63	-	7,522	-	-	-	-
$\Delta h_{\text{ср}}$, мм	-0,094	-0,077	-0,110	-0,857	-10,95	19,08	12,033	-47,04	0,910
K	0,278	0,770	-0,214	0,166	1,108	6,547	2,843	1,388	0,389
$Q_{\text{сл}}$, т	0,235	0,517	0,047	0,241	1,630	11,47	4,916	2,253	0,332
$F_{\text{сл}}$, мм^2	$0,88\cdot$ $\cdot 10^{-5}$	$0,13\cdot$ $\cdot 10^{-4}$	$0,45\cdot$ $\cdot 10^{-5}$	0,521	3,999	25,36	10,946	8,134	0,365
$F_{\text{пр}}$, мм^2	$-0,3\cdot$ $\cdot 10^{-4}$	$-0,19\cdot$ $\cdot 10^{-4}$	$-0,40\cdot$ $\cdot 10^{-4}$	-0,640	-5,462	17,61	8,463	-13,54	0,645
n_{K1}	0,082	0,231	-0,068	0,161	1,068	2,825	1,462	1,329	0,731
n_{K2}	0,086	0,325	0,154	0,106	0,700	1,735	0,929	0,810	0,753
n_{K3}	0,216	0,434	-0,001	0,285	1,949	6,534	3,110	2,836	0,627
n_{K4}	0,291	0,604	-0,023	0,267	1,818	8,873	3,936	2,600	0,462
S=0,604 R=0,944									

Первое приближение запишется следующим образом (см. также таблицу):

$$N = 13,81 - 0,094\Delta h_{\text{ср}} - 0,3\cdot 10^{-4}F_{\text{пр}} + 0,88\cdot 10^{-5}F_{\text{сл}} + 0,216n_{K3} + \\ +0,291n_{K4} + 0,235Q_{\text{сл}} + 0,278K + 0,082n_{K1} + 0,086n_{K2}.$$

Таблица 2. Второе приближение

N	14,66	14,84	14,48	-	7,985	-	-	-	-
$\Delta h_{\text{ср}}$, мм	-0,094	-0,077	-0,110	-0,855	-11,05	19,19	12,13	47,03	0,911
K	0,231	0,704	-0,243	-0,141	0,954	5,454	2,376	1,155	0,402
$Q_{\text{сл}}$, т	0,206	0,478	-0,065	0,217	1,493	10,14	4,355	1,999	0,343
$F_{\text{сл}}$, мм^2	$0,94\cdot$ $\cdot 10^{-5}$	$0,135\cdot$ $\cdot 10^{-4}$	$0,53\cdot$ $\cdot 10^{-5}$	0,556	4,498	27,30	11,83	9,665	0,380
$F_{\text{пр}}$, мм^2	$-0,31\cdot$ $\cdot 10^{-4}$	$-0,21\cdot$ $\cdot 10^{-4}$	$-0,41\cdot$ $\cdot 10^{-4}$	-0,667	-5,981	17,84	8,773	15,59	0,682
n_{K3}	0,202	0,386	0,016	0,303	2,132	5,618	2,913	3,195	0,732
n_{K4}	0,249	0,551	-0,053	0,234	1,617	7,622	3,397	2,215	0,476
S=0,615 R=0,942									

Таблица 3. Третье приближение

N	15,38	15,56	15,20	-	8,377
$\Delta h_{\text{ср}}$, мм	-0,95	- 0,079	- 0,112	-0,861	-11,64
$Q_{\text{сл}}$, т	0,229	0,496	-0,037	0,241	1,684
$F_{\text{сл}}$, мм^2	$0,93\cdot$ $\cdot 10^{-5}$	$0,134\cdot$ $\cdot 10^{-5}$	$0,52\cdot$ $\cdot 10^{-5}$	0,547	4,429
$F_{\text{пр}}$, мм^2	$-0,32\cdot$ $\cdot 10^{-5}$	$-0,23\cdot$ $\cdot 10^{-5}$	$-0,42\cdot$ $\cdot 10^{-5}$	-0,698	-6,969
n_{K3}	0,249	0,407	0,091	0,414	3,089
n_{K4}	0,367	0,547	0,188	0,509	3,998
S=0,621 R=0,941					

Доверительные интервалы коэффициентов регрессии следующих параметров проходят через нуль: K, $Q_{\text{сл}}$, n_{K1}, n_{K2}, n_{K3}, n_{K4}. По критерию II отбрасывается только n_{K2} по критерию III должно отбрасываться n_{K1} и n_{K2} а по критерию IV – n_{K1}, n_{K2}, K. Так как в программе многофакторного анализа сначала работает критерий III, то второе приближение, после отбрасывания n_{K1} и n_{K2} будет следующим:

$$N = 14,66 - 0,094\Delta h_{\text{ср}} - 0,31\cdot 10^{-4}F_{\text{пр}} + 0,94\cdot 10^{-5}F_{\text{сл}} + 0,202n_{K3} + \\ +0,249n_{K4} + 0,206Q_{\text{сл}} + 0,231K.$$

При этом коэффициент множественной регрессии уменьшается с $0,944$ до $0,942$, а остаточное среднеквадратическое отклонение увеличивается с $0,604$ до $0,615$.

Во втором приближении доверительные интервалы коэффициентов $K\,Q_{\text{сл}}$, n_{K4} регрессии параметров проходят через нуль, но критерии II и III дают отрицательный ответ на отбрасывание этих независимых параметров. По критерию IV можно отбросить независимый параметр K. Тогда в третьем приближении уравнение регрессии запишется следующим образом:

$$N = 15,38 - 0,095\Delta h_{\text{ср}} - 0,32\cdot 10^{-4}F_{\text{пр}} + 0,93\cdot 10^{-5}F_{\text{сл}} + 0,367n_{K4} + \\ +0,249n_{K3} + 0,229Q_{\text{сл}}.$$

Хотя после третьего приближения доверительный интервал у независимого параметра $Q_{\text{сл}}$ проходит через нуль, но ни один из выше рассмотренных критериев (II, III, IV) не отбрасывает его, так как t_{iI}, равный $1,684$, достаточно близок к $t_\alpha = 1,96$ и влияние на зависимую переменную N значительно ($r_{14} = 0,241$).

После третьего приближения коэффициент множественной корреляции равен $0,941$, а остаточное среднеквадратическое отклонение равно $0,621$.

Отбрасывание неэависимых параметров n_{K1}, n_{K2} согласуется с физическим процессом прокатки на блюминге, потому что среднее обжатие до первого ящичного калибра (где прокатка идет со стесненным уширением) определяется только размерами слитка и его положением при начальном обжатии, а в дальнейшем – размерами поперечного сечения конечного раската, последовательностью расположения и шириной калибров, которые в какой-то степени характеризуются параметрами n_{K1} и n_{K2}. Отсюда очевидно, что и количество кантовок мало влияет на среднее обжатие за цикл прокатки.

Заключение

Таким образом, предлагаемые критерии ранжирования независимых переменных и отбрасывания несущественных параметров, проверенные на многочисленных примерах из более чем 30-летней практики успешного проведения статистических исследований в машиностроении, металлургии и медицине [4, 5], которые подтверждают их надёжность и эффективность для проведения разнообразных многофакторных статистических исследований.

Список литературы

1. *Реноме* Статистические методы в алгоритмах и примерах (из практики прокатного производства). Учебное пособие ". ISBN 978-5-98947-081-5„СПб.: "Реноме", декабрь 2007, - 182с.
2. *Von H. Knüppel , Stumpf A., Wiezorke B.* Mathematische Statistik in Eisenhüttenwerken, Archive fürdas Eisenhüttenwesen, №8, 1958. Перевод № 1492. НИИТЯЖМАШ Уралмашзавода, 1968.
3. *Логоватовский А.А.* Нормирование процессов на блюминге. М.: Металлургия, 1966. 220с.
4. *Коцарь С.Л., Поляков Б.Н., Макаров Ю.Д., Чичигин В.А.* Статистический анализ и математическое моделирование блюминга М: Металлургия, 1974. 280с.
5. *Поляков Б.Н.* Повышение качества технологий, несущей способности конструкций, долговечности оборудования и эффективности автоматических систем прокатных станов. - СПб.: Реноме , 2006, - 528с.

Статья поступила в редакцию 25.10.2008