ПРАКТИЧЕСКИЙ АНАЛИЗ

Конкурирующие регрессии: критерии и процедуры отбора[1)]

Ершов Э.Б.

Предложен и реализован новый подход к определению наборов факторов для регрессии при заданном множестве потенциальных аргументов и фиксированной выборке. Выбираются наборы, являющиеся для используемых критериев качества регрессий конкурирующими, и такие, что не отвергается нормальность ошибок. К искомым регрессиям предъявляется требование быть гармоничными, обобщающее предложенное Хелвигом понятие коинцидентности. Несуществование одновременно конкурирующих, нормальных и гармоничных регрессий (KNH-регрессий) в условиях доверия к предположениям МНК интерпретируется как наличие в выборке нетипичных наблюдений. Предложен класс процедур «регрессионного тримминга», выявляющих и корректирующих такие наблюдения с целью нахождения скорректированных KNH-регрессий. Приведены примеры, использующие данные из классических работ по регрессионному анализу.

1. Введение

В эконометрике известны так называемые проблемные задачи, для которых отсутствуют строгие постановки, позволяющие получать однозначно интерпретируемые и признаваемые в общем случае решения. К числу таких ставших уже классическими задач, безусловно, относится задача выбора множества факторов для линейного регрессионного уравнения при заданном наборе потенциальных факторов и фиксированной объясняемой переменной.

Возможные теоретические подходы к выбору таких множеств факторов и практически применяемые методы их нахождения рассматривались в научной и учебной литературе многократно и со многих позиций, в том числе в специальных обзорных работах [14, 19, 25, 27, 32, 36, 39, 45, 63], монографиях общего характера [1, 2, 3, 7, 8, 15, 59, 61, 62] и оригинальных статьях, часть которых включена в список литературы к данной работе.

[1)] Статья представляет собой расширенный вариант доклада с тем же названием на VII Международной школе-семинаре «Многомерный статистический анализ и эконометрика» (Республика Армения, поселок Цахкадзор, 21–30 сентября 2008 г.), организаторами которой являлись ЦЭМИ РАН, Московская школа экономики МГУ, Российско-Армянский (Славянский) государственный университет, Армянский государственный экономический университет и журнал «Прикладная эконометрика».

Ершов Э.Б. – к.э.н., профессор кафедры математической экономики и эконометрики ГУ ВШЭ.

Статья поступила в Редакцию в октябре 2008 г.

Исходные предположения, на которых основываются разнообразные предлагаемые методы выбора предпочтительного уравнения, как правило, принимаются в качестве аксиом и лишь в редких случаях могут тестироваться с использованием данных выборки ограниченного и тем более малого объема.

Общий вывод, к которому можно придти, непредвзято анализируя содержание упомянутых публикаций, состоит в том, что применяемые в совокупности методы из множества потенциально возможных регрессий только выделяют подмножество конкурирующих между собой регрессий, каждая из которых является оптимальной или субоптимальной для одного или нескольких количественных, скалярных критериев качества регрессии. Поскольку такие критерии в рамках эконометрической теории представляют собой случайные величины-статистики, свойства которых определяются неизвестным в общем случае характером генеральной совокупности изучаемых величин и характером используемой выборки, то у исследователя имеется возможность субъективного выбора аксиом, относящихся к используемым данным.

Исследователь, выбирая аксиомы, может и даже должен принимать во внимание цели, для достижения которых будет применяться искомая регрессия. Но при этом возникает опасность того, что выбираемые так исходные предположения не будут согласованы с фактическим, труднотестируемым характером данных. В этих условиях естественно попытаться выделить из конкурирующих регрессий некоторое множество-ядро, которому по предположению, принадлежат «наилучшие» регрессии, какими бы методами они не определялись. Из такого понимания задачи следует, что к регрессиям из ядра должны предъявляться качественные требования, не противоречащие предположениям-аксиомам, на которых базируется применение частных критериев качества регрессий, и в то же время достаточно общие для того, чтобы в ядро не включались «спорные» конкурирующие регрессии, для которых есть основания сомневаться в оправдываемости «критических» аксиом, признаваемых обязательными к выполнению.

Предлагается практически реализуемый подход к определению такого ядра. Множество определений наилучших регрессий может пополняться за счет вводимых понятий, определений и конструкций. Возникает возможность проверки того, будут ли новые определения приводить к регрессиям из определяемого таким образом ядра. Выбор регрессии из ядра или даже конструирование зависимости объясняемой переменной от факторов с использованием регрессий из ядра как самостоятельных объектов представляет собой важную, но самостоятельную задачу, которой предполагается посвятить отдельное исследование.

Рассматривается следующая, не имеющая общепринятой постановки и метода решения проблема: для заданной величины y из множества факторов $\Omega_M(x_1, x_2, \ldots, x_M)$ выбрать подмножество $\omega_m = (x_{j(1)}, x_{j(2)}, \ldots, x_{j(m)})$, определяющее «наилучшее» уравнение $y = a_0 + \sum_j x_j a_j + e,\ j \in \omega_m$, оцениваемое методом наименьших квадратов (МНК) по данным n-мерной выборки $(y_k; x_{k1}, \ldots, x_{kM})$, $k = 1, \ldots, n$. Для нее предлагается относительно новая постановка и подход к решению, состоящий из нескольких этапов, на которых выявляются и используются специфические особенности данных. При этом будем исходить из признаваемой специалистами целесообразности различать задачу определения наилучшей регрессии для заданной выборки и задачу отбора множеств факторов, которые следует в первую очередь использовать в уравнениях для объясняемой переменной при значениях факторов, не включенных в вы-

борку, или при конструировании регрессии по данным других выборок, порождаемых, по предположению, общей для таких выборок и для имеющейся выборки генеральной совокупностью или моделью данных.

2. Множество конкурирующих наборов факторов

Многими авторами были предложены и применяются при выборе множества факторов ω_m различные «критерии качества регрессий» (ККР).

Для большинства критериев известны предположения, при которых они имеют теоретическое обоснование. В дальнейшем будем использовать традиционные для эконометрики обозначения:

$$X(\omega) = (i_n, x_{j(1)},\ldots, x_{j(m)}) \equiv X,$$

где $i_n' = (1,...,1)$ – n-мерный вектор;

$\hat{a}(\omega_m) = (X'X)^{-1}X'y$ – МНК-оценка вектора коэффициентов

$a' = (a_0, a_{j(1)},\ldots,a_{j(m)})$,

$P(\omega_m) \equiv X(X'X)^{-1}X'$,

$\hat{e}(\omega_m) = (I_n - P(\omega_m))y$,

$RSS(\omega_m) = \hat{e}(\omega_m)'\hat{e}(\omega_m)$,

$R^2(\omega_m) = 1 - RSS(\omega_m) / \sum_k (y_k - \bar{y})^2$.

Будем предполагать, что матрица $X(\Omega_M)'X(\Omega_M)$ невырождена и, следовательно, при заданном множестве потенциальных факторов Ω_M метод наименьших квадратов реализуем для любого подмножества факторов ω_m при $1 \leqslant m \leqslant M$.

Кратко охарактеризуем критерии, которые будут вычисляться на множестве всевозможных регрессий при фиксированных y, $\Omega_M(x)$ и выборке с тем, чтобы определить регрессии, являющиеся претендентами на роль наилучшей регрессии. Предлагаемая схема выделения множества таких регрессий инвариантна по отношению к выбору используемых скалярных ККР.

Основными максимзируемыми критериями являются:

$\bar{R}^2 \equiv R^2_{adj} = 1 - (n-1)(1-R^2)/(n-m-1)$ [61, 62];

$\check{R}^2 = 1 - n(1-R^2)/(n-m-1)$ [7, 24, 63];

$\tilde{R}^2 = 1 - \{(n-3)/(n-m-1)\}(1-R^2)F(1;1;0{,}5(n-m+1);1-R^2)$,

где $F(\alpha;\beta;\gamma;z)$ – специальная гипергеометрическая функция [46], и для математического ожидания $E\tilde{R}^2$ статистики $\tilde{R}^2(\omega_m)$ при $n > (m+1) \geqslant 3$ имеем $E\tilde{R}^2(\omega_m) = \eta^2$ – коэффициент детерминации для регрессии с набором факторов ω_m для $(m+1)$-мерной нормальной случайной величины

$$(y; x_{j(1)},\ldots,x_{j(m)}) = (y;\ \omega_m),\ x_j \in \omega_m;$$

$$\tilde{\tilde{R}}^2 = R^2 - (m-2)(1-R^2)/(n-m-1) - 2(n-3)(1-R^2)^2/\{(n-m-1)(n-m+1)\},$$

где $\tilde{\tilde{R}}^2$ – статистика, аппроксимирующая статистику $\tilde{R}^2$ при больших n [6];

$$R^2_{\min} = R^2 - \{8m(n-m-1)/[(n-1)(n+1)]\}^{0,5}(1-R^2),$$

где $R^2_{\min}$ – «нижняя доверительная граница» для $\mathcal{R}^2$ [2];

$$\mathrm{HELL} = \sum_i \tilde{r}(x_i; y)^2 \left\{ \sum_j \left| \tilde{r}(x_i; x_j) \right| \right\}^{-1},$$

где $x_i, x_j \in \omega_m$, $\tilde{r}(x, v)$ – выборочное значение коэффициента корреляции для переменных x, v [30, 31].

В качестве минимизируемых критериев часто используются:

$AIC = \ln(RSS/n) + 2(m+1)/n$ [12];

$BIC = \ln(RSS/n) + \{(m+1)/n\}\ln(n)$ [53];

$J \equiv PC = (n+m+1)\, RSS/(n-m-1)$ [14, 32, 43, 50];

$FPE = J/n$ [11];

$H = RSS/(n-m-1)$ [61, 62];

$HOCK \equiv S_p = RSS(\omega)/\{(n-m-1)(n-m-3)\}$ [20, 32, 63];

$СКОП = RSS(\omega)/\{(n-m-1)(n-m-2)\}$ – критерий «среднеквадратическая ошибка прогнозирования» [1, 4, 17, 54, 59, 64];

$PRESS = \sum_k \hat{e}_k^2 / (1 - P_{kk})^2$ [13, 52, 60];

$SRSS = \sum_k \hat{e}_k^2 / (1 - P_{kk})$ [32, 52];

$SHOCK = \{RSS(\omega_m) + \mathrm{RSS}(\Omega_M)\}/\{(n-m-1)(n-m-3)\}$ [32, 63];

$MAL \equiv C_p = RSS(\omega_m)/\tilde{\sigma}_e^2(\Omega_M) + 2(m+1) - n$, $MALL = MAL - (m+1)$ [28, 43, 44];

$MOO = \max_k |\hat{e}_k / y_k| \cdot 100$; $COO = n^{-1} \sum_k \left| \hat{e}_k / y_k \right| \cdot 100$ (для числа коэффициентов в регрессии используется обозначение $m + 1 \equiv p$).

Множественность критериев и трудности тестирования альтернативных теоретических условий корректного применения отдельных критериев с использованием данных выборки ограниченного объема делают обоснованный выбор из числа реализуемых ККР практически нерешаемой задачей. Обращает на себя внимание то, что некоторые критерии, имеющие свои обоснования, отличаются друг от друга лишь сомножителями, зависящими только от параметров n, m и констант 1, 2, 3. Взаимоотношения между наборами факторов, выбираемыми с использованием различных критериев при фиксированных объясняемой переменной, множестве потенциальных факторов Ω_M и выборке, а также ранги наборов при их упорядочивании по значениям ККР изучены недостаточно. Фактически проанализированы взаимосвязи критериев C_p, S_p, J [35], известны результаты, относящиеся к асимптотической эквивалентности некоторых критериев (см., например, [38, 58]), и арифметические соотношения, следующие непосредственно из определений критериев, представляющих собой функции от аргументов (n), (m) и RSS.

В статистических и эконометрических пакетах программ представлены только наиболее простые ККР. Даже в случае, когда методы тестирования условий при-

менимости таких критериев известны, их автоматическое тестирование не предусматривается и почти всегда не выполняется. Косвенные признаки невыполнения таких условий, в том числе связанные со спецификой используемой выборки, часто игнорируются, и, как следствие, особенности выборки не выявляются и не учитываются при наивно доверчивом выборе множества факторов.

Уровень изученности проблемы выбора применяемого критерия качества регрессии иллюстрирует ситуация с «несмещенным» критерием $\tilde{R}^2$. До последнего времени этот критерий не использовался из-за отсутствия эффектовного метода вычислений его значений. Такой метод был предложен и реализован в работе [5]. Заметим, что гипотеза нормальности генеральной совокупности для случайной величины $(y;\omega_m)$, на которой базируется применение ряда критериев, в том числе и $\tilde{R}^2$, тестируется в этой работе с помощью статистики W Шапиро – Уилка [7, 22, 23, 34, 55, 56, 57], которая фактически «вытеснена» из эконометрических пакетов и учебников асимптотической статистикой JB [18]. Статистика W вычисляется для каждой из переменных $y, x \in \omega_m$ и их ортогональных комбинаций, представленных в виде главных компонент их выборочной ковариационной матрицы. Этот прием известен, по-видимому, давно [49] и почти очевиден, но в статистической практике не используется, что возможно и даже скорее всего объясняется опасением столкнуться с отклонением гипотезы нормальности.

Доказано [5], что независимо от результатов тестирования этой гипотезы при $R^2 < 1$ выполняются неравенства $\tilde{R}^2 < \min\left(R^2_{adj}; \tilde{\tilde{R}}^2\right) \le \max\left(R^2_{adj}; \tilde{\tilde{R}}^2\right) < R^2$. Следовательно, если гипотеза нормальности не отвергается, то статистики R^2_{adj}, $\tilde{\tilde{R}}^2$ и R^2 можно считать положительно смещенными относительно $\mathfrak{R}^2$. Но значения статистик R^2_{adj} и $\tilde{\tilde{R}}^2$ не упорядочиваются, т.е. при фиксированных n и m в зависимости от значения R^2 выполняются неравенства $R^2_{adj} > \tilde{\tilde{R}}^2$ и $R^2_{adj} < \tilde{\tilde{R}}^2$. Для этих неравенств и для уравнения $R^2_{adj} = \tilde{\tilde{R}}^2$ найдены общие решения – множества троек (n,m,R^2). Для критерия $\tilde{\tilde{R}}^2$ получены формулы и разработана программа, вычисляющая его значения для любых n и m при $n > m + 1 \ge 3$. Известно [6], что статистика $\tilde{\tilde{R}}^2$ представляет собой функцию от полной системы достаточных статистик и единственную функцию от статистики R^2, чье математическое ожидание равно детерминированной величине $\mathfrak{R}^2$.

В охарактеризованных условиях, когда выбор какого-либо одного критерия качества регрессии затруднен, предлагается вычислять значения всех или нескольких отобранных эвристически или алгоритмически из охарактеризованных критериев. Эта рекомендация близка к позиции Себера, выраженной следующим образом: «Из приведенного рассмотрения ясно, что выбор критерия во многом зависит от того, каким образом модель собираются использовать. Поскольку очевидно, что дальнейшее исследование требует определенных свойств различных мер, то при сравнении моделей рекомендуется всегда вычислять не одну, а несколько мер» [7, с. 360]. Но

причины, из-за которых следует характеризовать набор факторов и регрессию значениями многих критериев, а также не ограничиваться сравнением этих критериев для разных регрессий, по нашему мнению и как будет видно из дальнейшего, не сводятся только к различиям направлений использования оцениваемых зависимостей.

Совместно с Н.А. Толмачевой разработана программа, вычисляющая значения задаваемых ККР для всех вариантов множеств факторов (при $M \leq 10$) и для каждого такого критерия определяющая его экстремальное (наибольшее или наименьшее) значение и множество факторов, для которого оно достигается, а также наборы факторов с близкими к экстремальным значениями выбранных критериев (для разных определений «ближайших» регрессий).

Многочисленные экспериментальные расчеты подтвердили предположение, согласно которому, как правило, для критерия $K_\varphi(\omega)$ существуют такие наборы факторов $\omega^s, s = 1\ldots,S$, что $K_\varphi(\omega^s) \approx \max(\min)_\omega K_\varphi(\omega) \equiv K_\varphi(\omega^{opt})$. Это позволяет определить $\Omega(K_\varphi)$ – *множество конкурирующих* (для критерия K_φ) *наборов факторов*. Тогда для критериев $K_\varphi(\omega)$, $\varphi \in \Psi$, где Ψ – множество отобранных ККР, определяется *множество конкурирующих* (для Ψ) *наборов факторов* $\Omega(K) \equiv \bigcup_\varphi \Omega(K_\varphi) \equiv \Omega(K[\Psi])$.

Такие наборы факторов рассматривались многими исследователями в иллюстративных примерах для эмпирически выбираемых в процессе анализа критериев (см., например, [3, гл. 6; 7, гл. 12]). Себер отмечает, что в пакете BMDP для каждого из трех критериев R^2, R^2_{adj}, $MAL \equiv C_p$ находится задаваемое число субоптимальных наборов факторов.

Дрейпер и Смит так характеризуют практическую неразрешимость задачи универсального выбора критерия качества регрессий и одновременно метода определения «наилучшего» набора факторов: «Для реализации такого выбора нет однозначной статистической процедуры» (с. 9); «Чтобы окончательно выбрать модель, требуются дополнительные априорные соображения и здравый смысл экспериментатора» (с. 28); «Никакой метод не будет хорошо работать при всех условиях, как бы хорошо он не проявил себя на частном примере» (с. 58).

При практической трудности классификации и тестирования достаточных условий применения отдельных критериев для всевозможных вариантов регрессий предлагается на первом этапе решения поставленной задачи в качестве промежуточного результата рассматривать именно множество конкурирующих наборов факторов $\Omega(K[\Psi])$. На втором этапе к *конкурирующим* регрессиям с $\omega \in \Omega(K[\Psi]) \equiv \Omega(K)$ предъявим качественные «дополнительные априорные» требования, в которых, по нашему мнению, проявляется «здравый смысл экспериментатора».

3. Конкурирующие нормальные и гармоничные регрессии

Наиболее простым и логичным качественным требованием к набору факторов, для которого регрессия претендует на роль «наилучшей» регрессии, является, конечно, его принадлежность к множеству конкурирующих наборов. При сравнении двух таких «вложенных» ($\omega^1 \subset \omega^2$) или «невложенных» ($\omega^1 \not\subset \omega^2 \not\subset \omega^1$) наборов ω^1 и ω^2

с целью определить из них более предпочтительный традиционно в прикладной эконометрике принимается гипотеза нормальности ошибок для искомой «истинной» регрессии. Все критерии качества регрессий на стадии их теоретической интерпретации также в той или иной степени базируются на этой гипотезе. Поэтому естественно требовать от рассматриваемых регрессий, и в первую очередь от конкурирующих, чтобы для них *не отвергалась гипотеза нормальности ошибок*. Такие регрессии и их наборы факторов будем называть *нормальными*, вводя для множества нормальных наборов факторов и соответствующих регрессий обозначение $\Omega(N)$.

Гипотезу нормальности набора факторов $\omega_m \in \Omega(N)$ можно тестировать многими способами. Проведенный анализ соответствующих публикаций показал, что предпочтение можно отдать использованию уже упоминавшейся статистики Шапиро – Уилка $W_n(e)$, рассчитываемой по МНК-остаткам $\hat{e}_k(\omega_m)$, $k = 1,\ldots,n$. Гипотеза нормальности ошибок ($e_к$) не отвергается, если $W_n(\hat{e}) > w_{n,p}$, где $w_{n,p}$ – получаемое из таблицы критическое значение статистики W_n и p – задаваемая доверительная вероятность (уровень значимости). Величины $w_{n,p}$ известны по крайней мере для $n = 3,\ldots,50$ и $p = 0{,}01;\ 0{,}02;\ 0{,}05;\ 0{,}10;\ 0{,}50;\ 0{,}90;\ 0{,}95;\ 0{,}98;\ 0{,}99$. В тесте Шапиро – Уилка используются уровни значимости $p = 0{,}5$ и $p > 0{,}5$. Принадлежность $W_n(\hat{e})$ к интервалу между критическими значениями $w_{n,p}$ позволяет получить представление о степени оправдываемости гипотезы нормальности. Тест Шапиро – Уилка включен в статистический пакет SPSS и в оригинальные программы, разработанные Толмачевой Н.А. в сотрудничестве с автором статьи.

Таким образом определяется *множество конкурирующих и нормальных наборов факторов* $\Omega(KN) \equiv \Omega(K) \cap \Omega(N)$.

Второе качественное требование к наборам факторов $\omega_m \subset \Omega_M$ определим как требование *гармоничности* набора и соответствующей регрессии. Оно состоит в том, что знаки МНК-оценок $\hat{a}_j$ коэффициентов a_j при факторах $x_j \in \omega_m$ должны совпадать со знаками выборочных коэффициентов корреляции $\tilde{r}(y;x_j)$ и ковариации $\tilde{cov}(y;x_j)$ для объясняемой переменной y и факторов x_j.

Для того чтобы это определение можно было распространить на случай, когда $\hat{a}_j = 0$ для фактора $x_j \in \omega_m$, требование гармоничности сформулируем в более общем виде, а именно как систему неравенств $\hat{a}_j \tilde{r}(y;x_j) \geq 0$, $x_j \in \omega_m$. Тот факт, что аналогичные неравенства можно считать выполняющимися для всех факторов, которые не включены в набор ω_m, не влияет на свойства гармоничных регрессий и возможности их использования при решении рассматриваемой проблемной задачи. Регрессии, для которых выполняются неравенства $\hat{a}_j \tilde{r}(y;x_j) > 0$, $x_j \in \omega_m$, можно называть *вполне гармоничными*.

Определение гармоничных регрессий целесообразно дополнить требованием надежного определения знака произведения статистик $\hat{a}_j$ и $\tilde{r}(y;x_j)$ или $\tilde{cov}(y;x_j)$. Его можно было бы интерпретировать как условие неотклонения нелинейной гипотезы неотрицательности или положительности произведений $a_j r(y;x_j)$ при $x_j \in \omega_m$, в которых ненаблюдаемые теоретические величины a_j и $r(y;x_j)$ должны быть определены с учетом особенностей данных и предполагаемых использований регрессий.

Эта гипотеза должна тестироваться с использованием только выборочных значений коэффициентов корреляции и МНК-оценок коэффициентов a_j. Но метод, позволяющий оценивать вероятность выполнения системы неравенств $a_j r(y;x_j) \geq 0$, $x_j \in \omega_m$, даже при упрощающих предположениях о регрессии с множеством факторов ω_m и о случайной величине $(y;\omega_m)$, не разработан. В этих условиях приходится ограничиваться эвристическими оценками вероятностей выполнения неравенств $a_j^{ист} \geq 0$, $a_j^{ист} \leq 0$, $r(y;x_j) \geq 0$, $r(y;x_j) \leq 0$, $a_j r(y;x_j) \geq 0$ в предположениях о нормальности ошибок и случайной величины $(y;\omega_m)$ или о детерминированности значений факторов. Если получаемые оценки вероятностей выполнения неравенств $a_j^{ист} r\left(y;x_j\right) \geq 0$ достаточно велики, то будем считать, что регрессия признается гармоничной «надежно». Задача строго обоснованного тестирования гипотезы гармоничности при различных упрощающих, но реалистичных предположениях, безусловно, заслуживает внимания. В данной статье ограничимся простым определением гармоничности, использующим статистики $\hat{a}_j \tilde{r}(y;x_j)$.

Но возможен и иной подход к объяснению того, почему по отношению к конкурирующим регрессиям предъявляется требование быть гармоничными. Можно рассматривать задачу оценки или даже вычисления вероятности того, что регрессия с детерминированными факторами $x_j \in \omega_m$, независимыми нормальными ошибками с дисперсией σ_e^2 и известными истинными значениями коэффициентов $a_j^{ист}$ будет при оценивании методом наименьших квадратов признана гармоничной, т.е. будут выполняться неравенства $\hat{a}_j \tilde{r}(y;x_j) \geq 0$, $x_j \in \omega_m$. Эта задача относительно легко решается в простейшем случае, когда имеется всего два фактора, которые можно считать центрированными и нормированными.

Охарактеризуем метод решения этой задачи. Пусть значения переменных x_1, x_2 в выборке образуют столбцы матрицы (x). Тогда находится множество $\Gamma(a^{ист};r_{12})$ значений нормально распределенных МНК-оценок $\hat{a}$ коэффициентов при факторах $\{\hat{a} \sim N(a^{ист};\sigma_e^2 (x'x)^{-1})\}$ такое, что при $\hat{a} \in \Gamma(a^{ист};r_{12})$ оцененная регрессия признается гармоничной. При этом корреляционная матрица (r) для факторов определяется известным значением коэффициента корреляции $r_{12} = r(x_1;x_2)$, а неравенства $r(y;x_1) \geq (\leq) 0$, $r(y;x_2) \geq (\leq) 0$ эквивалентны неравенствам $\hat{a}_1 + r_{12}\hat{a}_2 \geq (\leq) 0$, $r_{12}\hat{a}_1 + \hat{a}_2 \geq (\leq) 0$. Интегрируя по имеющему простую структуру множеству $\Gamma(a^{ист};r_{12})$ нормальную плотность для двумерного вектора $\hat{a}$, можно вычислить искомую вероятность $P(a^{ист}, r_{12}, \sigma_e^2)$ как функцию четырех аргументов. Предположение о том, что однофакторные и гармоничные регрессии с аргументами x_1 и x_2 являются конкурирующими, будем интерпретировать в виде приближенного равенства коэффициентов детерминации для этих регрессий: $\tilde{r}(y;x_1)^2 \cong \tilde{r}(y;x_2)^2$. Получаемое таким образом решение подтверждает предположение, согласно которому вероятность $P(a^{ист}, r_{12}, \sigma_e^2)$, как правило, будет значительно превосходить вероятность получить негармоничную регрессию, что и является аргументом, мотивирующим формулирование требования гармоничности. Задачу вычисления вероятности P предполагается рассмотреть в отдельной публикации.

Очевидно, что выделять гармоничные и особенно «надежно гармоничные» наборы факторов целесообразно из множества конкурирующих и нормальных наборов $\Omega(KN)$. Для множества гармоничных или кратко H-регрессий (H~harmonic) и наборов факторов введем обозначение $\Omega(H)$. Таким образом определены множества $\Omega(KH)$, $\Omega(NH)$, $\Omega(KNH)$, представляющие собой результаты операций пересечения множеств, включенных в их обозначения

$$\{\text{def: } \Omega(AB) \equiv \Omega(A) \cap \Omega(B), \Omega(ABC) \equiv \Omega(A) \cap \Omega(B) \cap \Omega(C)\}.$$

Наибольший интерес представляет множество наборов факторов $\Omega(KNH)$, называемое *ядром конкурирующих, нормальных, гармоничных регрессий* и являющееся результатом второго этапа решения задачи. Но следует иметь в виду, что введенные множества наборов факторов в конкретных случаях могут быть пусты. Так, для фиксированной выборки *возможно*: $\Omega(KNH) = \varnothing$ и даже $\Omega(H) = \varnothing$, но всегда существуют конкурирующие регрессии ($\Omega(K) \neq \varnothing$) и гармоничные регрессии ($\Omega(H) \neq \varnothing$), так как однофакторная регрессия является гармоничной ($\omega_1(x_j) \in \Omega(H)$) и даже вполне гармоничной, если $\tilde{r}(y;x_j) \neq 0$. Поэтому результатом второго этапа может быть и обнаружение несовместности требований конкурентности, нормальности, гармоничности наборов факторов и регрессий.

Поясним мотивы, по которым к конкурирующим регрессиям целесообразно предъявлять требование быть гармоничными. Гармоничные регрессии обладают следующими представляющими интерес для теоретических и прикладных исследований свойствами.

Во-первых, определение *гармоничности* инвариантно относительно невырожденных, линейных, сепарабельных преобразований переменных. Поэтому такие регрессии удобно рассматривать в центрированных и нормированных переменных, для которых знаки МНК-оценок $\hat{\beta}_j$ коэффициентов β_j при факторах совпадают со знаками оценок $\hat{a}_j$ при $x_j \in \omega_m$.

Во-вторых, в так называемом факторном разложении коэффициента детерминации $R^2(\omega) \equiv \sum_{j \in \omega} \tilde{r}(y;x_j)\hat{\beta}_j(\omega)$ вклады факторов x_j неотрицательны, а для вполне гармоничных регрессий даже положительны, т.е. $\tilde{r}(y;x_j)\hat{\beta}_j(\omega) > 0$, $x_j \in \omega$. Целесообразность и обоснованность включения в регрессию факторов с отрицательнами вкладами в показатель-статистику R^2, по мнению автора статьи, нуждается в специальном мотивировании. Если для вполне гармоничной регрессии вклады факторов принадлежат полуоткрытому интервалу $(0; R^2]$, то для негармоничных регрессий вклады некоторых факторов отрицательны и даже могут быть меньше, чем $(-R^2)$. Такой набор факторов и объясняемую переменную можно охарактеризовать как внутренне противоречивую, несамосогласованную с позиций регрессионного анализа совокупность переменных. Условие неотрицательности вкладов факторов в R^2 можно принимать в качестве определения гармоничных наборов факторов и регрессий.

В-третьих, множество гармоничных регрессий при заданных переменных y и $x_j \in \Omega_M$ обладает свойством, формулируемым в виде следующего утверждения, непосредственно связанного с часто применяемыми для выбора наилучшей регрессии «методом исключения» факторов и «шаговым регрессионным методом» пополнения множества факторов.

Для $\omega_m \equiv (x_1,\ldots,x_m) \in \Omega(H)$ существует такая последовательность факторов $\{x_{j(1)},\ldots,x_{j(m)}\}$, $x_{j(s)} \in \omega_m$, что гармоничны все «вложенные» регрессии с наборами факторов $\omega^1 = (x_{j(1)})$, $\omega^2 = (x_{j(1)},x_{j(2)}),\ldots,$ $\omega^{m-1} = (x_{j(1)},\ldots,x_{j(m-1)})$, $\omega^m = (x_{j(1)},\ldots,x_{j(m)}) \equiv \omega_m$, отличающиеся одним исключаемым или включаемым фактором.

Доказательство. В работах [37, 65] рассматривались достаточные и необходимые условия, при которых МНК-оценки коэффициентов $\hat{a}_j(\omega_{m-1})$ и $\hat{a}_j(\omega_m)$ при факторе x_j в регрессиях с наборами аргументов ω_{m-1} и $\omega_m \equiv (\omega_{m-1};x_i)$, где x_i – исключаемый из ω_m фактор, имеют одинаковые или противоположные знаки. В частности, было показано, что $\hat{a}(\omega_{m-1})\hat{a}_j(\omega_m) \geq 0$ при $j \neq i$, если модули t-статистик $t_i(\omega_m)$, $t_j(\omega_m)$ для коэффициентов при факторах x_i и x_j в регрессии с набором факторов ω_m удовлетворяют неравенству $|t_i(\omega_m)| \leq |t_j(\omega_m)|$. Следовательно, исключая из гармоничного набора факторов ω_m любую переменную x_i с наименьшим значением модуля t-статистики, получаем гармоничную регрессию. С регрессией, имеющей набор факторов ω_{m-1}, поступаем таким же образом и т.д. Поскольку регрессия с одним фактором всегда гармонична, получаем искомую последовательность регрессий.

Таким образом, гармоничные регрессии представляют собой последовательности вложенных гармоничных регрессий, отличающихся одним фактором. Это представление не обязано быть единственным. Среди гармоничных регрессий выделяются финально-гармоничные регрессии с наборами факторов ω_m, для которых при заданном множестве потенциальных факторов Ω_M не существуют «включающие» их гармоничные регрессии с факторами $x_j \in \omega_{m+h}$, $h > 0$ и $\omega_m \subset \omega_{m+h}$.

В частном случае гармоничные регрессии изучались Хеллвигом (Zdislav H. Hellwig) как регрессии, «*обладающие свойством коинцидентности*», у которых отсутствует «эффект катализа» [30, 31]. Хеллвиг так определил это свойство: при условиях, что факторы положительно коррелированны с переменной y, т.е. $\tilde{r}_{0j} \equiv \tilde{r}(y;x_j) > 0$ (условие всегда выполнимо с помощью перехода от переменной x_j к $-x_j$, если $\tilde{r}_{0j} \neq 0$, $j = 1,\ldots,m$), и $|\tilde{r}_{0j}| \neq 1$, выполняются неравенства:

1) $\tilde{r}_{0i} \cdot \tilde{r}_{0j} > \tilde{r}_{ij} \equiv \tilde{r}(x_i;x_j) > 0$, в которых $i \neq j$;
2) $\hat{a}_i > 0$, $i = 1,\ldots,m$.

Если отказаться от условий (1) и от неравенств $\tilde{r}_{0j} \neq +1,-1$, а неравенства (2) трансформировать в $\hat{a}_j \geq 0$, то получаем определение гармоничной регрессии. Элжбета Максимиак [41] нашла два варианта условий, достаточных для того, чтобы регрессия обладала свойством коинцидентности. Хеллвиг и Максимиак рассматрива-

ли это свойство как присущее «хорошим» регрессиям. Следует, однако, заметить, что свойство коинцидентности также определялось в терминах выборочных значений соответствующих статистик, т.е. без внимания к их случайному характеру.

Эффект катализа проявляется в том, что выборочное значение $\tilde{r}(y;x_j)$ коэффициента корреляции имеет знак, отличный от знака коэффициента при факторе x_i в предполагаемом существующем, истинном уравнении регрессии. Поскольку такое уравнение неизвестно (неизвестен даже набор его факторов) и ищется, то, заменяя понятие коинцидентности более общим понятием гармоничности, можно исходить из следующего объяснения причин возникновения нежелательного или требующего содержательного обоснования эффекта катализа: негармоничность регрессии, если она «надежно» выявляется при использовании имеющихся данных (следует помнить, что знаки МНК-оценок $\hat{a}_i$ и коэффициентов $\tilde{r}_{0j}$ не обязательно надежно определяются), возможно является следствием неправильного выбора множества факторов или наличия в данных нетипичных наблюдений. Тогда негармоничность конкурирующей регрессии может рассматриваться как предупреждение о выборе и использовании набора факторов, нуждающегося в дополнительном оправдании, и гармоничность искомых конкурирующих и нормальных регрессий представляется качественно правдоподобным и естественным постулатом.

Примеры негармоничных регрессий, признаваемых удовлетворительно моделирующими зависимость объясняемой переменной от факторов и даже наилучшими, относительно редко, но все же встречаются в серьезных публикациях. Так, например, в [6] вычисление МНК-оценок коэффициентов иллюстрируется примерами 27.1 и 27.2, в которых регрессии с полными наборами факторов ($\omega_m = \omega_M \equiv \Omega_M$) не являются гармоничными. Но в этих примерах, заимствуемых из работ других авторов, выбор множества потенциальных факторов не ставится под сомнение, нормальность данных и ошибок для регрессии не тестируется, «надежность» определения знаков выборочных коэффициентов корреляции $\tilde{r}(y;x_j)$ и МНК-оценок $\hat{a}_j$ коэффициентов a_j не анализируется. Фактически принимается без обсуждения и проверки гипотеза однородности выборки или по крайней мере отсутствия в ней нетипичных наблюдений. Последнее допущение оправданно при разработке теории, но вызывает опасения, когда в конкретных ситуациях обнаруживаются эффект катализа, маскирующие связи переменных и требующие разъяснений знаки оценок коэффициентов при факторах.

4. Выявление и учет нетипичных наблюдений с помощью процедур регрессионного тримминга

Гипотезы нормальности и гармоничности конкурирующих регрессий могут отклоняться по отдельности или противоречить друг другу по многим причинам. Среди них могут быть и такие, что задача конструирования линейного уравнения связи объясняемой переменной с факторами из выбранного множества в виде оцениваемой методом наименьших квадратов регрессии должна признаваться не соответствующей исследуемому объекту.

Будем предполагать, что рассматриваемая задача все же сформулирована адекватно и что причины отклонения этих гипотез следует искать в особенностях используемой выборки, проявляющихся в наличии в ней нетипичных наблюдений,

которые не были выявлены на стадии предварительного анализа данных. Более сложный случай, когда выборка фактически состоит из нескольких, не разделяемых очевидным образом подвыборок, для которых искомая зависимость должна представляться регрессиями с разными значениями коэффициентов при факторах, различающимися дисперсиями ошибок или даже своими наборами факторов, здесь не рассматривается и заслуживает специального исследования.

Выявлять, корректировать, не удаляя, и учитывать при оценивании конкурирующих регрессий нетипичные наблюдения предлагается на третьем этапе решения задачи, используя процедуру регрессионного тримминга или, коротко, RTR-процедуру. В названии этой процедуры используется информативный в данном случае англоязычный термин «trimming» – удаление заусенцев, выглаживание, отделка.

На этом этапе для множества факторов ω рассматривается регрессия

$$y = a_0 + \sum_{j\in\omega} x_j a_j + vb + u \qquad (1)$$

с независимыми ошибками $u = (u_k)$ и $v = (v_k)$ такими, что $u \sim \mathrm{N}(0; \sigma_u^2 I_n)$ и $v_k = -1, 0$ и $+1$ с вероятностями, равными соответственно (p), $(1 - 2p)$, (p) или даже p_-, $\mathrm{p}_0 = (1 - \mathrm{p}_- - p_+)$, p_+. Параметры $\pi \equiv (a, b, \sigma_u^2, p)$ оцениваются методом максимального правдоподобия (МП) на подмножестве $\Pi(\omega)$ возможных вариантов ошибок $v = (v_k)$ с вариантами $B_r(v) \in \Pi(\omega)$, $r = 1,\ldots,n$. Общее число возможных вариантов велико и равно $(3^n - 3)$. Поэтому задача оценивания параметров, предствляющая собой в данном случае задачу комбинированного, т.е. непрерывного и дискретного математического программирования, рассматривается на множестве $\Pi(\omega)$, состоящем всего из n элементов-вариантов ошибок (v_k). Задача оценивания регрессии (1), тестирования ее нормальности и гармоничности (по переменным $x_j \in \omega$) решается для каждого такого варианта $B_r(v)$.

Исходная гипотеза, относящаяся к механизму «засорения» значений переменной y ошибками (v_k), имеющими дискретное распределение и приводящими к появлению нетипичных наблюдений, состоит в том, что такие ошибки проявляются в наблюдениях с «большими» модулями МНК-остатков $\hat{e}_k$ для простейшей, исходной регрессии

$$y = a_0 + \sum_{j\in\omega} x_j a_j + e \equiv X(\omega)a + e \qquad (2)$$

или большими модулями нормализованных остатков $\check{e}_k \equiv \hat{e}_k/(1 - P_{kk})^{0,5}$, имеющих равные дисперсии, если ошибок (v_k) в регрессии действительно нет.

Заметим, что критерий SPSS определяется в виде суммы квадратов остатков $\check{e}_k$, а критерий PRESS – в виде суммы квадратов иначе нормализованных остатков $\hat{e}_k/(1 - P_{kk}) \equiv \tilde{e}_k$. Для того чтобы определить варианты дамми-переменной v в регрессии (1), наблюдения будем упорядочивать одним из трех способов: в соответствии с невозрастанием модулей остатков $\hat{e}_k$, $\check{e}_k$ или $\tilde{e}_k$, т.е. так, что $|\hat{e}_{k(1)}| \geq |\hat{e}_{k(2)}| \geq \ldots \geq |\hat{e}_{k(n)}|$ или что аналогичные неравенства выполняются для какого-либо из вариантов нормализованных остатков.

Тогда вариант $B_r(v)$ дамми-переменной $v = (v_k)$ определим следующим образом: $v_{k(s)} = \text{sign}(\hat{e}_{k(s)})$ при $1 \le s \le r$, $v_{k(q)} = 0$ при $r + 1 \le q \le n$. Для нормализованных остатков значения переменной v в варианте $B_r(v)$ определяются аналогичным образом. Для ненулевых v_k можно использовать и определение $v_k = \text{sign}(-\hat{e}_k)$, что для регрессии (1) приводит лишь к изменению знака оценки коэффициента b. Таким образом, вариант $B_r(v)$ характеризуется числом r ненулевых величин v_k и включаемых в регрессию возможных корректировок $v_k b$ значений объясняемой переменной.

Для регрессии с вариантом $B_r(v)$ ошибок $v = (v_k)$: находятся МП-оценки параметров π, включая t-статистики для оценок $\hat{a}_j$ и остатки $\tilde{u} = (\tilde{u}_k)$; вычисляются значения критериев $K_\varphi \in K[\psi] \in \text{K}[\Psi]$, где ψ – выбираемое подмножество критериев ($\psi \subset \Psi$); вычисляется значение критерия KML, минимизация которого эквивалентна максимизации критерия метода максимального правдоподобия; с использованием статистики $W(\tilde{u}\{B_r\})$ Шапиро – Уилка тестируется нормальность ошибок $\tilde{u} = (\tilde{u}_k)$; проверяется гармоничность скорректированной регрессии $y - v(B_r)\hat{b} = Xa + u$. Для минимизируемого критерия KML получено явное выражение

$$\text{KML} = \ln(1 - R^2) - 2\{(r_-/n)\ln(r_-/n) + (r_0/n)\ln(r_0/n) + (r_+/n)\ln(r_+/n)\}, \quad (3)$$

в котором используются МП-оценки вероятностей ошибок $v_k = -1,0,+1$: $p = r/2n$ или $p_{-1} = r_-/n$, $p_0 = r_0/n$ и $p_{+1} = r_+/n$, где r_-, r_0, r_+ – числа отрицательных, нулевых и положительных величин v_k для варианта ошибок $B_r(v)$.

Полученная в результате обширная информация, относящаяся к каждому из n вариантов регрессии (1), анализируется с целью выделить наборы факторов ω_m, признаваемые одновременно конкурирующими, нормальными, гармоничными и имеющими коэффициенты a_j при факторах $x_j \in \omega_m$, для которых гипотезы $a_j = 0$ отвергаются. При этом используется выявляемая нормальность регрессий. В разработанной версии программного комплекса, включающего процедуры всех трех этапов решения рассматриваемой задачи, анализ результатов третьего этапа, т.е. регрессионного тримминга, реализована в двух вариантах: предназначенном для углубленного анализа, когда регрессии (2) оцениваются для всех n вариантов $B_r(v)$, $r = 1,\ldots,n$, переменной $v = (v_k)$; предназначенном для практического анализа, когда регрессии (2) оцениваются при возрастающих значениях r ($r = 1,\ldots,q$) до получения скорректированной регрессии, признаваемой нормальной или нормальной и гармоничной.

Процедура RTR применяется к набору факторов ω из ядра $\Omega(KNH)$, из $\Omega(KH/N)$ и из $\Omega(KN/H)$ [def: $\omega \in \Omega(AB/C) \sim \omega \in \Omega(A) \cap \Omega(B)$, но $\omega \notin \Omega(C)$], что сокращает число рассматриваемых на этом этапе наборов факторов. Из наборов факторов $\omega \in \Omega(KH/N)$, не являющихся по определению этого множества нормальными, выделяются такие наборы $\omega \in \Omega R(KNH/N)$, для которых регрессия (1) с вариантом ошибок $B_r(v)$ с возможно наименьшим числом r ненулевых элементов v_k признается удовлетворяющей предъявляемым требованиям. Таким же образом из множества $\Omega(KN/H)$ конкурирующих и нормальных, но негармоничных регрессий выделяется его подмноже-

ство $\Omega R(KNH/H)$ регрессий, признаваемых гармоничными в результате применения тримминга, корректирующего нетипичные наблюдения.

В результате на этом этапе определяется множество *устойчиво (или робастно) конкурирующих (относительно RTR), нормальных и гармоничных* RKNH-регрессий: $\omega \in \Omega R(KNH) = \Omega(KNH) \cup [\Omega R(KNH/N) \cup \Omega R(KNH/H)]$.

Если множество-ядро $\Omega R(KNH)$ оказывается пустым, а такие случаи встречались при проведении экспериментальных расчетов, использующих данные из работ других исследователй, то в регрессию (1) включаются две дамми-переменные со значениями 0 и −1 или 0 и +1 в наблюдениях, номера которых определяются по величинам остатков $\hat{e}_k$, $\check{e}_k$ или $\tilde{e}_k$ для рассматриваемых наборов факторов. Эти дамми-переменные в регрессии включаются с оцениваемыми коэффициентами b_- и b_+. Для таких регрессий выполяется анализ нормальности и гармоничности. Такое выявление и учет нетипичных наблюдений позволяют, как правило, определить непустое множество $\Omega R(KNH)$, т.е. найти решение задачи в предлагаемой постановке.

Число наборов факторов, включенных в $\Omega R(KNH)$, можно сократить, используя методы парных сравнений нормальных регрессий. Для пар вложенных наборов факторов из ядра $\Omega R(KNH)$, используя нормальность ошибок, можно проверить гипотезу равенства нулю коэффициентов при переменных, которыми такие наборы различаются. Для пары невложенных регрессий можно тестировать гипотезу «предпочтительности» одной из них с помощью статистики Вуонга [29, 66], базирующейся на информационном критерии KLIC Кульбака − Лейблера.

Наборы факторов для регрессии из ядра $\Omega R(KNH)$ или часть таких наборов, отобранных в результате парных сравнений регрессий и образующих, по определению, *множество-ядро конкурентных наборов факторов* $\Omega R(KNH)$, можно рассматривать в качестве опорных элементов, с использованием которых будут конструироваться функциональные зависимости переменной (y) от факторов из Ω_M по данным используемой выборки и других выборок для выбранной модели данных. Эта также проблемная задача, требующая корректной формулировки, в настоящее время исследуется.

5. Два обозримых, иллюстрирующих примера

Предложенные процедуры применялись к широко используемым в учебных целях, содержащимся в монографиях, статьях и учебниках по прикладной статистике, многомерному статистическому анализу и эконометрике [2, 3, 6, 7, 8, 9, 29, 40, 63] примерам с различными числами потенциальных факторов (M) и наблюдений (n), а также к линейным по параметрам регрессионным зависимостям, оцениваемым по данным российских таблиц « Затраты − Выпуск » и национальных счетов.

Предлагаемый подход к выделению ядер $\Omega R(KNH)$ конкурирующих и $\Omega R(KNH)$ конкурентных регрессий продемонстрируем на двух обозримых примерах с относительно небольшими значениями M и n.

Пример 1. Регрессии для количества тепла, выделяемого при производстве цемента.

В этом широко используемом, детально разбираемом рядом авторов примере [3, 7, 8] $M = 4, n = 13$, факторы сильно коррелированны, нормальность генеральной со-

вокупности уверенно отвергается. Для всех $\omega \in \Omega_M$ рассчитаны значения выбранных критериев $K_\varphi \in K(\Psi)$, протестирована нормальность ошибок и проверена гармоничность регрессий (см. табл. 5.1, 5.2). В этих таблицах и далее наборы факторов приводятся в упрощенном виде («12» ~ (x_1,x_2)),курсивом выделяются признаваемые конкурирующими наборы. Принадлежность регрессии к множествам нормальных (N), гармоничних (H) и KNH-регрессий отмечается знаком «+» в соответствующем столбце таблицы.

В табл. 5.3 приведены критические значения статистики Шапиро − Уилка для чисел наблюдений $n = 13$ и $n = 20$, используемые в примерах 1 и 2.

Конкурирующими, нормальными и гармоничными (*KNH-регрессиями*) признаны регрессии с наборами факторов (x_1,x_4) и $(x_1,x_2,x_4) \equiv (1,2,4)$. Для этих регрессий и других *конкурирующих регрессий (1,2), (1,2,3), (1,3,4), (1,2,3,4)* проверяется их принадлежность к *устойчиво KNH-регрессиям*. В табл. 5.4 приведены значения статистики W для исходных регрессий и результаты тестировния нормальности и гармоничности RTR-скорректированных регрессий. Тестировалась нормальность ошибок (u_k) и проверялась гармоничность (по $x_j \in \omega$) скорректированных регрессий $y = (Xa + vb) + u$ с одной дамми-переменной $v = (v_k)$, определяемой по остаткам $\hat{e}_k$. «Спорные» регрессии с факторами (1,2) и (1,2,3) оказались устойчиво (относительно RTR) ненормальными или негармоничными; для регрессии (1,4) подтверждена нормальность (хотя и с понижением уровня нормальности: W = 0,967) и гармоничность; для регрессии (1,3,4) выявлена RTR-нормальность (для скорректированной регрессии W = 0,957), а для регрессии (1,2,3,4) − RTR-гармоничность при сохранении нормальности. «Идеальная» регрессия $\omega(1,2,4) \in \Omega(KNH)$ является RTR-устойчиво нормальной (W = 0,948) и RTR-гармоничной.

Таблица 5.1.

Значения максимизируемых критериев $K_\varphi \cdot 10^5$

$j \in \omega$	R^2	R^2_{adj}	$R^2_{\min}$	$\tilde{\tilde{R}}^2$	$\tilde{R}^2$	HELL	N	H	KNH
1	53395	49158	39421	-	-	53395	+	+	
2	66627	63593	56620	-	-	66627		+	
3	28587	22095	7175	-	-	28587		+	
4	67454	64495	57696	-	-	67454		+	
1,2	*97867*	***97441***	***96841***	***97860,26***	***97860,21***	***97691***		+	
1,3	54817	45780	33051	51414,12	50770,98	44943		+	
1,4	*97247*	*96697*	*95921*	*97234,48*	*97234,37*	*97033*	+	+	+
2,3	84703	81643	77333	84312,52	84294,42	83577			
2,4	68006	61607	52594	66300,00	66122,19	67959		+	
3,4	93529	92235	90412	93459,18	93457,85	93286		+	
1,2,3	*98228*	***97638***	***97058***	***98025,29***	***98025,26***	*89282*	+		
1,2,4	*98234*	***97645***	***97067***	***98030,97***	***98030,94***	***96894***	+	+	+
1,3,4	*98128*	***97504***	***96891***	***97913,04***	***97913,00***	*94128*		+	
2,3,4	97282	96376	95486	96965,07	96964,95	89688			
1,2,3,4	*98238*	***97356***	***96728***	***97789,19***	***97789,14***	***96049***	+		

Таблица 5.2.

Значения минимизируемых критериев K_φ

$j\in\omega$	AIC	BIC	SRSS	PRESS	HOCK	MOO	COO	N	H	KNH
1	4,886	4,973	1456,9	1699,6	11,506	20,46	9,30	+	+	
2	4,552	4,639	1040,4	1202,1	8,239	18,45	6,82		+	
3	5,313	5,400	2248,2	2616,4	17,631	30,79	11,94		+	
4	4,527	4,614	1023,5	1194,2	8,035	17,36	7,43		+	
1,2	***1,955***	***2,086***	***72,8***	***93,9***	***0,643***	***3,90***	***2,04***		+	
1,3	5,009	5,139	1598,5	2218,1	13,634	17,99	9,58		+	
1,4	*2,211*	*2,341*	*94,1*	*121,2*	*0,831*	*6,93*	*2,22*	+	+	+
2,3	3,926	4,056	537,1	701,7	4,616	11,64	4,45			
2,4	4,664	4,794	1122,2	1461,8	9,654	16,40	6,98		+	
3,4	3,066	3,196	226,0	294,0	1,953	7,12	2,96		+	
1,2,3	***1,924***	***2,098***	***65,0***	***90,0***	***0,668***	***4,49***	***1,71***	**+**		
1,2,4	***1,921***	***2,095***	***63,4***	***85,4***	***0,666***	***4,26***	***1,73***	+	+	+
1,3,4	***1,979***	***2,153***	***68,2***	***94,5***	***0,707***	***4,04***	***1,82***		+	
2,3,4	2,352	2,526	101,6	146,9	1,025	4,70	2,03	**+**		
1,2,3,4	*2,073*	*2,290*	***71,0***	*110,3*	*0,855*	***4,38***	***1,72***	**+**		

Таблица 5.3.

Критические значения статистик W_{13} и W_{20} Шапиро − Уилка

p	0,01	0,02	0,05	0,10	0,50	0,90	0,95	0,98	0,99
$w_{13,p}$	0,814	0,837	0,866	0,889	0,945	0,974	0,979	0,984	0,986
$w_{20,p}$	0,868	0,884	0,905	0,920	0,959	0,979	0,983	0,986	0,988

Таблица 5.4.

Результаты анализа RTR-устойчивости множеств ω

$j\in\omega$	K	N	H	KNH	$w_{13,p'}$	$< W \leq$	$w_{13,p''}$	RTR-N	RTR-H	$\omega\in\Omega R(KNH)$
1,2	+		+		$0{,}889_{0,1}$	0,905	$0{,}945_{0,50}$		+	
1,4	+	+	+	+	$0{,}974_{0,9}$	0,975	$0{,}979_{0,95}$	+	+	+
1,2,3	+	+			$0{,}974_{0,9}$	0,977	$0{,}979_{0,95}$	+		
1,2,4	+	+	+	+	$0{,}945_{0,5}$	0,964	$0{,}974_{0,90}$	+	+	+
1,3,4	+		+		$0{,}889_{0,1}$	0,944	$0{,}945_{0,50}$	+	+	+
1,2,3,4	+	+			$0{,}945_{0,5}$	0,970	$0{,}974_{0,90}$	+	+	+

Для регрессий (1,4), (1,2,4), (1,3,4) и (1,2,3,4) необходимо выяснить, не приводит ли RTR-учет нетипичных наблюдений к такому изменению значений контрольных критериев R^2_{adj} (заменяет критерий RSS при данном ω), MOO и COO, использованных при определении конкурирующих регрессий, если их вычислять по остаткам в скорректированных регрессиях, что среди отобранных регрессий выявится явный претендент на роль наилучшей регрессии либо регрессии, уступающие явным образом другим регрессиям.

Для признаваемых нормальными регрессий с вложенными наборами факторов $(x_1,x_4) \subset (x_1,x_2,x_4)$ и $(x_1,x_4) \subset (x_1,x_3,x_4)$ отвергаются гипотезы $a_2 = 0$ и $a_3 = 0$, так как в нормальных и гармоничных регрессиях с различными вариантами переменной v модули соответствующих t-статистик достаточно велики ($|t_2| \in [5{,}6; 8{,}5]$, $|t_3| \in [4{,}9; 7{,}1]$). Но регрессия (x_1,x_4) проигрывает регрессиям с этими трехфакторными наборами по значениям контрольных критериев R^2_{adj}, AIC, BIC, SRSS, PRESS, HOCK, MOO и COO, хотя по другим критериям такие преимущества не выявляются. Поэтому набор (x_1,x_4) признан «неконкурентным» по отношению к наборам (x_1,x_2,x_4) и (x_1,x_3,x_4). В табл. 5.5 приведены значения некоторых контрольных критериев, рассчитываемых для регрессии ω по остаткам $\hat{e}$ в уравнении $y = Xa + e$, по остаткам $\hat{u}$ в уравнении $y = (Xa + v(r)b) + u$ и по «остаткам» $\hat{g} = \hat{u} + (r)\hat{b}$.

Для регрессий (1,3,4) и (1,2,3,4), являющихся претендентами на включение в ядро конкурентных регрессий $\Omega R(KNH)$, переход от остатков $\hat{e}$ в исходной регрессии к остаткам $\hat{g}$ для RTR-регрессий мало изменяет значения максимизируемого критерия R^2_{adj} (в сторону уменьшения) и минимизируемого критерия COO (в сторону увеличения), но даже приводит к уменьшению значения критерия MOO. В то же время для регрессий-претендентов значения критериев близки к их значениям для «идеальной» регрессии с $(x_1,x_2,x_4) \in \Omega(KNH)$.

Таблица 5.5.

Значения контрольных критериев для регрессий-кандидатов на включение в множество устойчиво KNH-регрессий, $\omega \in \Omega R(KNH)$

Остатки $j \in \omega$	Критерии								
	R^2_{adj}			MOO			COO		
	$\hat{e}$	$\hat{u}$	$\hat{g}$	$\hat{e}$	$\hat{u}$	$\hat{g}$	$\hat{e}$	$\hat{u}$	$\hat{g}$
1,4	0,967	0,996	0,961	6,93	1,86	5,74	2,22	0,73	2,44
1,2,4	0,976	0,998	0,971	4,26	0,96	3,86	1,73	0,48	1,82
1,3,4	0,975	0,997	0,970	4,04	1,30	3,94	1,82	0,53	1,84
1,2,3,4	0,974	0,998	0,966	4,38	1,15	3,64	1,72	0,46	1,82

Из табл. 5.5 следует, что регрессия (1,4) проигрывает другим регрессиям по значениям критериев (хотя проигрыш по R^2_{adj} незначителен) и поэтому может рассматриваться как неконкурирующая с ними. Следовательно, *регрессии (1,2,4), (1,3,4) и (1,2,3,4) образуют ядро устойчиво конкурирующих, нормальных и гармоничных регрессий.* Из них только $\omega(x_1,x_2,x_4) \in \Omega(KNH) \subset \Omega R(KNH)$.

Для пар отобранных вложенных регрессий с наборами факторов (1,2,4)⊂(1,2,3,4) и (1,3,4)⊂(1,2,3,4) протестированы гипотезы равенства нулю коэффициентов при переменных x_3 и x_2, т.е. $a_3(x_1,x_2,x_3,x_4;v) = 0$ и $a_2(x_1,x_2,x_3,x_4;v) = 0$. Эти гипотезы не были отвергнуты ($|t_2| \cong 1{,}36, |t_3| \cong 0{,}91$). Для невложенных регрессий с факторами $(x_1,x_2,x_4;v)$ и $(x_1,x_3,x_4;v)$ тест Вуонга показал их «равноправие» (эквивалентность), что не противоречит значениям критериев в табл. 5.5. Поэтому в ядре конкурентных наборов факторов ΩR(*KNH*) оставлены только наборы (x_1,x_2,x_4) и (x_1,x_3,x_4).

Детальный анализ «Примера Хальда», выполненный в [3, 7], но без тестирования нормальности ошибок, без проверки гармоничности регрессий и, конечно, без применения корректировки данных с использованием процедуры RTR, привел к рассмотрению конкурирующих наборов факторов (1,2), (1,4), (1,2,3), (1,2,4), (1,3,4) и к выбору регрессии с $\omega(x_1,x_2)$ как наиболее предпочтительной. Но этот набор факторов не был признан ни нормальным, ни нормальным после корректировки данных (RTR-нормальным), хотя гипотеза нормальности ошибок этими авторами фактически использовалась. Поэтому регрессия с факторами x_1, x_2 не включена в ядра наборов факторов Ω(*KNH*), Ω*R*(*KNH*) и ΩR(*KNH*).

Пример 2. Регрессии для урожайности зерновых в районах некоторой области.

Этот пример используется в работе [2] как сквозной пример, на котором иллюстрируются многие постановки и методы решения задач регрессионного анализа. Для него нормальность генеральной совокупности уверенно отвергается, факторы слабее коррелированны с объясняемой переменной, чем в примере 1. В исходных данных имеются 20 наблюдений и 5 факторов, но фактор x_1 был исключен, так как $r(x_1,x_3) \approx 0{,}98$, что позволило снизить опасность проявления мультиколлинеарности факторов. Оставшимся четырем факторам даны номера 1,...,4.

Была применена та же схема анализа множества возможных регрессий, что в примере 1. Это позволяет избежать подробного описания ее этапов и промежуточных результатов.

В табл. 5.6 и 5.7 приведены значения критериев для конкурирующих вариантов множеств факторов, результаты тестирования нормальности ошибок в исходных уравнениях и проверки их гармоничности.

Из этих таблиц следует, что среди конкурирующих регрессий нет так называемых KNH-регрессий, поскольку все эти регрессии не признаются нормальными. Ни одна из 31 возможных регрессий не является нормальной, и велики значения критериев MOO и COO. Это позволяет предполагать отсутствие в данных важных факторов или наличие в них нетипичных наблюдений. В такой ситуации применение процедуры регрессионного тримминга с целью выявления RKNH-регрессий вполне оправданно и даже представляется необходимым.

Таблица 5.6.

Значения максимизируемых критериев $K_\varphi \cdot 10^5$

$j \in \omega$	R^2	R^2_{adj}	$R^2_{\min}$	$\tilde{\tilde{R}}^2$	$\tilde{R}^2$	HELL	H
1,3	46196	**39866**	**29674**	**43148**	**42783**	**46085**	+
2,3	48237	**42147**	**32342**	**45416**	**45093**	**48146**	+
1,2,3	48386	38708	27092	42015	41635	**47374**	+
1,3,4	51346	**42223**	**31273**	**45510**	**45195**	35711	
2,3,4	49823	**40415**	**29122**	**43715**	**43367**	**39221**	
1,2,3,4	51730	38858	26712	42188	41819	38592	

Таблица 5.7.

Значения минимизируемых критериев K_φ

$j \in \omega$	AIC	BIC	SRSS	PRESS	HOCK	MOO	COO	H
1,3	**0,991**	**1,140**	46,28	**54,51**	**0,1467**	**25,34**	11,71	+
2,3	**0,952**	**1,101**	**43,54**	**49,93**	**0,1413**	28,35	**10,47**	+
1,2,3	1,049	1,248	45,48	55,52	**0,1505**	**27,71**	**10,63**	+
1,3,4	**0,990**	**1,189**	**43,51**	**53,63**	**0,1503**	**24,29**	10,97	
2,3,4	**1,021**	**1,220**	**43,78**	**52,42**	0,1550	28,38	**10,09**	
1,2,3,4	1,082	1,331	**45,38**	61,65	0,1704	**25,58**	**10,42**	

Была применена RTR-процедура выявления нетипичных наблюдений. В качестве таких наблюдений были выделены наблюдения с номерами $k = 7, 19, 20$. Включение в регрессии переменной $v = (v_k)$ такой, что $v_7 = v_{19} = -1$, $v_{20} = +1$, $v_k = 0$ при $k \neq 7, 19, 20$, привело к результатам, характеризуемым в столбцах табл. 5.8, относящихся к скорректированным регрессиям. Для регрессий с наборами факторов (1,3) и (1,3,4) нормальность ошибок отвергается при всех вариантах дамми-переменной v. Для регрессий с факторами (1,3,4), (2,3,4) и (1,2,3,4) введение такой переменной не позволило добиться их RTR-гармоничности.

Таблица 5.8.

Результаты анализа RTR-устойчивости множеств ω

$j \in \omega$	Для уравнения $y = Xa + e$			Для уравнения $y = Xa + u + vb$			RTR-N	RTR-H
	$w_{20,p'}$	$< W \leq$	$w_{20,p''}$	$w_{20,p'}$	$< W \leq$	$w_{20,p''}$		
1,3	$0{,}905_{0,05}$	0,913	$0{,}920_{0,10}$	$0{,}920_{0,1}$	0,928	$0{,}959_{0,5}$		+
2,3		0,862	$0{,}868_{0,01}$	$0{,}959_{0,5}$	0,975	$0{,}979_{0,9}$	+	+
1,2,3		0,865	$0{,}868_{0,01}$	$0{,}959_{0,5}$	0,970	$0{,}979_{0,9}$	+	+
1,3,4	$0{,}920_{0,10}$	0,923	$0{,}959_{0,50}$	$0{,}920_{0,1}$	0,944	$0{,}959_{0,5}$		
2,3,4	$0{,}868_{0,01}$	0,882	$0{,}884_{0,02}$	$0{,}988_{0,99}$	0,989	$1{,}0_{1,0}$	+	
1,2,3,4	$0{,}905_{0,05}$	0,918	$0{,}920_{0,10}$	$0{,}959_{0,5}$	0,971	$0{,}979_{0,9}$	+	

Таблица 5.9.

Значения контрольных критериев для регрессий-кандидатов на включение в множество устойчиво *R*KNH-регрессий

Остатки $j\in\omega$	Критерии								
	R^2_{adj}			МОО			СОО		
	$\hat{e}$	$\hat{u}$	$\hat{g}$	$\hat{e}$	$\hat{u}$	$\hat{g}$	$\hat{e}$	$\hat{u}$	$\hat{g}$
2,3	0,827	0,978	0,793	16,77	5,53	15,73	5,90	2,15	6,68
1,2,3	0,818	0,975	0,780	17,29	5,84	15,75	5,77	2,23	6,63

Ядро *R*KNH-регрессий образуют уравнения с наборами факторов $(x_2,x_3)\subset(x_1,x_2,x_3)$. Для этих регрессий рассчитаны значения контрольных критериев, которые приводятся в табл 5.9. Из табл. 5.8 и 5.9 следует, что для этих регрессий значения критериев и статистик *W* близки при небольшом преимуществе регрессии (x_2,x_3). Поскольку для этих регрессий нормальность ошибок не отвергается, протестируем гипотезу равенства нулю «истинного» коэффициента при переменной x_1 для скорректированной регрессии с $\omega=(x_1,x_2,x_3)$. МНК-оценки коэффициентов a_j, $j\in\omega$ и их *t*-статистики t_j приводятся в табл. 5.10.

Таблица 5.10.

Оценки коэффициентов при факторах для $\omega\in\Omega R(KNH)$

ω	$\hat{a}_1$	t_1	$\hat{a}_2$	t_2	$\hat{a}_3$	t_3
2,3	–	–	0,4257	16,4	2,980	14,1
1,2,3	2,6254	0,77	0,3845	6,7	3,048	13,7

Очевидно, что «наилучшей» регрессией должно быть признано уравнение с факторами x_2 и x_3, так как $t_1<1$ и, следовательно, $\Omega R(KNH)=(x_2,x_3)$. Этот вывод совпадает с рекомендацией С.А. Айвазяна и В.С. Мхитаряна [2], базировавшейся на использовании критерия $R^2_{\min}$, но здесь используется иная, более общая аргументация.

6. Общие выводы и рекомендации

Определение «наилучшей» регрессии с использованием единственного, выбираемого без обоснования критерия качества регресии, без тестирования гипотезы нормальности ошибок и других неявно принимаемых гипотез не имеет надежного оправдания и может приводить к «случайному», даже неверному выбору множества факторов. В общем случае некорректна трактовка выбранной таким способом «наилучшей» регрессии как регрессии, имеющей «правильные» знаки МНК-оценок коэффициентов при факторах. Если гипотеза нормальности генеральной совокупности данных уверенно не отвергается, то относительно просто выделяются факторы, для которых знаки коэффициентов корреляции $r(y;x_j)$ надежно определены. Априорные представления о знаках коэффициентов при факторах в искомом урав-

нении для объясняемой переменной y, по-видимому, должны совпадать с получаемыми эмпирически и не могут индуцироваться результатами оценок коэффициентов для регрессии с набором факторов, который не был корректно проанализирован на включение в него «чужих», не только «лишних» в классическом смысле факторов и на выполнение используемых предположений (например, о нормальности ошибок и отсутствии нетипичных наблюдений).

Если знаки коэффициентов корреляции $r(y;x_j)$ определены уверенно, то гипотеза гармоничности искомых (а не единственной, наилучшей) конкурирующих и нормальных регрессий представляется качественно оправдываемым требованием.

Предъявляемое к конкурирующим регрессиям требование быть устойчивыми по отношению к выявлению нетипичных наблюдений, нормальными и гармоничными достаточно естественно и может применяться при выборе факторов, т.е. реализуемо с использованием процедур регрессионного тримминга.

Целесообразно продолжить исследования в следующих направлениях.

При упрощающих предположениях о генеральной совокупности данных или модели данных, например, о нормальности случайных величин $(y;\omega_m)$ или (ω_m), нормальности ошибок и детерминированности факторов, полезно и интересно в теоретическом отношении разработать методы тестирования точно формулируемой нелинейной гипотезы гармоничности регрессии.

Необходимо исследовать процедуры выявления нетипичных наблюдений с помощью различных вариантов регрессионного тримминга, корректировки значений объясняемой переменной в таких наблюдениях и выделения ядер регрессий и наборов факторов $\Omega R(KNH)$ и $\Omega\mathrm{R}(KNH)$.

Для задачи конструирования линейной зависимости переменной y от факторов в виде функции от регрессий из $\Omega R(KNH)$ и $\Omega\mathrm{R}(KNH)$ должна быть предложена просто интерпретируемая и имеющая решение постановка.

* *
*

СПИСОК ЛИТЕРАТУРЫ

1. *Айвазян С.А., Енюков И.С., Мешалкин Л.Д.* Исследование зависимостей. М.: Финансы и статистика, 1985.

2. *Айвазян С.А., Мхитарян В.С.* Прикладная статистика и основы эконометрики. М.: ЮНИТИ, 1998.

3. *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ. Книга 2. М.: Финансы и статистика, 1987. (Перевод монографии: Draper N.R., Smith H. Applied Regression Analysis. John Wiley and Sons, 1981.)

4. *Енюков И.С.* Оценивание параметров и критерии отбора информативных переменных в линейных регрессионных моделях со случайными аргументами // 11 Всесоюзная науч.-техн. конференция «Применение многомерного статистического анализа в экономике и оценке качества продукции»: Тезисы доклада. Тарту, 1981. С. 214–218.

5. *Ершов Э.Б.* Выбор регрессии, максимизирующий несмещенную оценку коэффициента детерминации // Прикладная эконометрика. 2008. № 4.

6. *Кендалл М., Стьюарт А.* Статистические выводы и связи. М.: Наука, 1973. (Перевод монографии: Kendall M.G., Stuart A. The Advanced Theory of Statistics. Vol. 2. Inference and Relationship. London: Charles Griffin and Company Limited, 1969.)

7. *Себер Дж.* Линейный регрессионный анализ. М.: Мир, 1980. (Перевод монографии: Seber G.A.F. Linear Regression Analysis. Wiley Series in Probability and Statistics. John Wiley and Sons, 1977.)

8. *Хальд А.* Математическая статистика с техническими приложениями. М.: ИЛ, 1956. (Перевод монографии: Hald A. Statistical Theory with Engineering Application. John Wiley and Sons, 1952.)

9. *Abt K.* On the Identification of the Significant Independent Variables in Linear Models // Metrika. 1967. Vol. 12. P. 2–15.

10. *Aitkin M.A.* Simultaneous Inference and the Choice of Variable Subset in Multiple Regression // Technometrics. 1974. Vol. 16. P. 221–227.

11. *Akaike H.* Statistical Predictor Identification // Annals of the Institute of Statistical Mathematics. 1970. 22. P. 203–217.

12. *Akaike H.* Information Theory and an Extension of the Maximum Likelihood Principle // B. Petrov, F. Csake (eds.) Second International Symposium on Information Theory. Budapest: Akademiai Kiado, 1973. P. 257–281.

13. *Allen D.M.* Mean Square Error of Prediction as a Criterion for Selecting Variables // Technometrics. 1971. Vol. 13. P. 469–475.

14. *Amemiya T.* Selection of Regressors // International Economic Review. 1980. Vol. 21. P. 331–354.

15. *Amemiya T.* Advanced Econometrics. Cambridge: Harvard University Press, 1985.

16. *Anscombe F.J.* Topics in the Investigation of Linear Relations Fitted by the Method of Least Squares // Journal of the Royal Statistical Society. 1967. 29. P. 1–52.

17. *Bendel R.B., Afifi A.A.* Comparison of Stopping Rules in Forward «Stepwise Regression» // Journal of the American Statistical Associations. 1977. Vol. 72. P. 46–53.

18. *Bera A., Jarque C.* Efficient Test for Normality, Heteroscedasticity, and Serial Independence of Regression Residuals: Monte Carlo Evidence // Economic Letters. 1981. 7. P. 313–318.

19. *Boyce H.J., Farhi A., Weischedel R.* Optimal Subset Selection. Lecture Notes in Economics and Mathematical Systems / H.E. Beckman, H.P. Kunzi (eds.) N.Y.: Springer-Verlag, 1974.

20. *Breiman L., Freedman D.* How Many Variables should be Entered in a Regression Equation? // Journal of the American Statistical Associations. 1983. Vol. 78. № 381. P. 131–136.

21. *Chow G.C.* The Selection of Variables for Use in Prediction: A Generalization of Hotelling`s Solution // L.N. Klein, M. Nerlove, S.C. Tsiang (eds.) Quantitative Econometrics and Development. N.Y.: Academic Press, 1980. P. 105–114.

22. *Csorgo M., Seshardi V., Yalovsky M.* Some Exact Tests for Normality in the Presence of Unknown Parameters // Journal of the Royal Statistical Society. Series B (Methodological). 1973. 35. № 3. P. 507–522.

23. *Dyer A.R.* Comparisons of Tests for Normality with a Cautionary Note // Biometrika. 1974. Vol. 61. P. 185–189.

24. *Ezekiel M.* Methods of Correlation Analysis. N.Y.: Wiley, 1930.

25. *Gaver K.M., Geisel M.S.* Discriminating among Alternative Models: Bayesian and Non-Bayesian Methods // P. Zarembka (eds.) Frontiers in Econometrics. N.Y.: Academic Press, 1974. P. 48–80.

26. *Golan A.* A Simultaneouse Estimation and Variable Selection Rule // Journal of Econometrics. 2001. 101. P. 165–193.

27. *Golan A., Judge G.G., Miller D.* Maximum Entropy Econometrics: Robust Estimation with Limited Data. N.Y.: John Wiley and Sons, 1980.

28. *Gorman J.W., Toman R.J.* Selection of Variables for Fitting Equation to Data // Technometrics. 1966. Vol. 8. P. 27–51.

29. *Green W.H.* Econometric Analysis. 6th ed. Prentice Hall: Pearson Education, Inc., 2008.

30. *Hellwig Z.* Problem Optymalnego wyboru predyktant (A Problem of Optimal Choice of Predicands) // Przeglad Statystyczne. 1968. № 3–4.

31. *Hellwig Z.* Efect katalizy, jego wykrywanie i usuwanie (The Effect of Catalysis, its Detection and Elimination // Przeglad Statystyczny. 1977. № 2. P. 179–192.

32. *Hocking R.R.* The Analysis and Selection of Variables in Linear Regression // Biometrics. 1976. № 62. P. 1–49.

33. *Hotelling H.* The Selection of Varieties for Use in Prediction with Some Comments on the General Problem of Nuisance Parameters // Annals of Mathematical Statistics. 1940. Vol. 11. P. 271–283.

34. *Huang C.J., Bolch B.W.* On the Testing of Regression Disturbances for Normality // Journal of the American Statistical Associations. 1974. Vol. 69. № 346. P. 330–335.

35. *Kinal T., Lahiri K.* A Note on Selection of Regressors // International Economic Review. 1984. Vol. 25. № 3.

36. *Lavergne P.* Selection of Regressors in Econometrics: Parametric and Nonparametric methods // Econometric Reviews. 1998. 17. P. 227–273.

37. *Leamer E.* A Result of the Sign of Restricted Least-Squares Estimates // Journal of Econometrics. 1975. 3. P. 387–390.

38. *Lien D., Vuong Q.H.* Selecting the Best Linear Regression Model. A Classical Approach // Journal of Econometrics. 1987. 35. P. 3–23.

39. *Lindley D.V.* The Choice of Variables in Multiple Regression // Journal of the Royal Statistical Society. 1968. 30. P. 31–53.

40. *Maddala G.S.* Introduction to Econometrics. 2nd ed. N.Y.: Macmillan Publishing Company, 1992.

41. *Maksymiak E.* O wlasnosci koincydencji i efekcie katalizy dia modeli opisywanych przez pewne pary korelacyjne // Preglad. Statystyczny. 1986–1987. № 4. P. 353–360.

42. *Mallows C.L.* Chosing Variables in Linear Regression: A Graphical Aid. Presented at the Central Regional Mitting of the Institute of Mathematical Statistics. Manhattan, Kansas, 1964.

43. *Mallows C.L.* Choosing a Subset Regression. Presented at the Joint Statistical Meeting. Los Angeles, California, 1966.

44. *Mallows C.L.* Some Comments on C_p // Technometrics. 1973. Vol. 15. P. 661–675.

45. *Miller A.J.* Subset Selection in Regression. London: Chapman & Hall, 1980.

46. *Olkin I., Pratt J.W.* Unbiased Estimation of Certain Correlation Coefficients // Annals of the Institute of Statistical Mathematics. 1958. 29.

47. *Pesaran M.H.* On the General Problem of Model Selection // Review of Economic Studies. 1974. Vol. 41. P. 153–171.

48. *Pesaran M.H.* On the Comprehensive Method of Testing Non-Nested Regression Models // Journal of Econometrics. 1982. 18. P. 263–274.

49. *Putter J.* Orthonormal Bases of Error Spacts and their Use for Investigating the Normality and Variance of Residuals // Journal of the American Statistical Associations. 1967. Vol. 62. P. 1022–1036.

50. *Rothman D.* Letter to the Editor // Technometrics. 1968. Vol. 10. P. 432.

51. *Sawa T.* Information Criteria for Discriminating among Alternative Regression Models // Econometrica. 1978. Vol. 46. P. 1273–1291.

52. *Schmidt P.* Methods of Choosing among Alternative Linear Regression Models. Chapel Hill, North Carolina: University of North Carolina, 1973.

53. *Schwarz G.* Estimating the Dimension of a Model // Annals of Statistics. 1978. 6. P. 461–464.

54. *Sclove S.L.* Improved Estimation of Regression Parameters. Tech. Report № 125. Palo Alto, California: Dep. of Statist., Stanford University, 1967.

55. *Shapiro S.S., Francia R.S.* An Approximate Analysis of Variance Test for Normality // Journal of the American Statistical Associations. 1972. Vol. 67. P. 215–216.

56. *Shapiro S.S., Wilk M.B.* An Analysis-of-Variance Test for Normality (Complete Samples) // Biometrika. 1965. Vol. 52. № 3/4. P. 591–611.

57. *Shapiro S.S., Wilk M.B., Chen H.J.* A Comparative Study of Various Tests for Normality // Journal of the American Statistical Associations. 1968. Vol. 63. № 324. P. 1343–1372.

58. *Shibata R.* An Optimal Selection of Regression Variables // Biometrika. 1981. Vol. 68. № 1. P. 45–54.

59. *Stein C.* Multiple Regression // Contribution to Probability and Statistics: Essays in honor of Harold Hotelling. Palo Alto, California: Stanford University Press, 1960. P. 424–443.

60. *Stone M.* Cross-Validatory Choice and Assessment of Statistical Predictions // Journal of the Royal Statistical Society. 1974. 30. P. 111–147.

61. *Theil H.* Economic Forecasts and Policy. 2nd ed. Amsterdam: North-Holland, 1961.

62. *Theil H.* Principles of Econometrics. N.Y.: John Wiley and Sons, 1971.

63. *Thompson M.L.* Selection of Variables in Multiple Regression: Part 1. A review and evaluation. Part 11. Chosen procedures, computations and examples // International Statistical Review. 1978. Vol. 46. № 1, 2. P. 1–19, 129–146.

64. *Tukey J.W.* Discussion (of Anscombe [1967]) // Journal of the Royal Statistical Society. 1967. 29. P. 47–48.

65. *Visco I.* On Obtaining the Right Sign of a Coefficient Estimate by Omitting a Variable from the Regression // Journal of Econometrics. 1978. 7. P. 115–117.

66. *Vuong Q.H.* Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses // Econometrica. 1989. Vol. 57. P. 307–334.

67. *Zhang P.* On the Distributional Properties of Model Selection Criteria // Journal of the American Statistical Associations. 1992. Vol. 87. P. 732–737.

68. *Zheng X., Loh W.-Y.* Consistent Variable Selection in Linear Models // Journal of the American Statistical Associations. 1995. Vol. 90. P. 151–156.