УДК 519.21 : 519.24

В.В. Крохин, Н.О. Кузьменко

АВТОМАТИЗАЦИЯ ВЫБОРА ОПТИМАЛЬНОЙ МОДЕЛИ ЛИНЕЙНОЙ РЕГРЕССИИ

Аннотация. Проведен анализ эффективности выбора оптимальной модели множественной линейной регрессии (МЛР) различными методами. Исследование проводилось с помощью метода имитационного моделирования. Были проанализированы пять методов выбора оптимальной МЛР. Анализ проводился с помощью специально разработанного программного обеспечения с использованием MATLAB.

Ключевые слова: автоматизация, оптимальная модель, множественная линейная регрессия, критерий оптимальности, имитационное моделирование, распараллеливание вычислительных процессов.

Введение. Выбор оптимальной модели регрессии базируется, прежде всего, на понимании исследователем механизмов порождающих имеющиеся данные, т.е. на использовании априорной информации. Вместе с тем можно сформулировать ряд критериев, позволяющих выбирать оптимальную модель, если класс, к которому принадлежит правильная модель, определяется основными принципами, которые нужно учитывать при выборе модели.

Под выбором оптимальной модели подразумевается, как это принято в регрессионном анализе, определение подмножества регрессоров из совокупности всех возможных независимых переменных, которое наилучшим образом «объясняет» наблюдаемые значения зависимой переменной.

Постановка задачи. В данной статье рассматривается ряд критериев, позволяющих выбрать оптимальную модель регрессии в классе линейных по параметрам моделей, а также проводится сравнительный анализ качества этих критериев путём имитационного моделирования. С целью автоматизации выбора оптимальной модели МЛР в среде МАТЛАБ был разработан программный продукт.

Решение задачи. Модель линейной регрессии (МЛР) многих переменных представляется следующим образом:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + U_i,\ i=1,2,\ldots,n, \quad (1)$$

где i обозначает номер наблюдения; X_{ji}, $j=1,\ldots,k$ - независимые (объясняющие, экзогенные, предикторные) переменные или регрессоры; Y - зависимая (объясняемая, эндогенная) переменная; U - возмущение (ошибка модели); n – объём выборки.

Модель (1) удобно записывать в матричной форме:

$$Y = X\beta + U, \quad (2)$$

где Y - наблюдаемый $[n\times 1]$ вектор значений зависимой переменной; X - $[n\times(k+1)]$ матрица, в которой первый столбец состоит из единиц, а остальные столбцы представляют наблюдаемые значения зависимых переменных; β является неизвестным $[(k+1)\times 1]$ вектором параметров (коэффициентов) модели и U является ненаблюдаемым $[n\times 1]$ вектором возмущений (ошибок модели).

В данной работе считается, что модель регрессии удовлетворяет условиям Гаусса – Маркова [1,4], а ошибки модели имеют нормальное распределение.

Обозначим через $\hat{\beta}$ оценки β, полученные по выборке. Тогда мы можем оценить ошибки модели u следующим образом:

$$\hat{U}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - \ldots - \hat{\beta}_k X_{ki},\ i = 1,\ldots,n. \quad (3)$$

Величины $\hat{U}_i$ называются выборочными ошибками или остатками. Чем меньше будут выборочные ошибки, тем большая часть изменения зависимой переменной Y обуславливается изменениями в независимых переменных. В матричных обозначениях формула (3) может быть записана в виде:

$$\hat{U} = Y - X\hat{\beta}. \quad (4)$$

Величины

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_k X_{ki}, \quad i = 1,2,\ldots,n \quad (5)$$

являются оценками значений зависимой переменной Y, полученными с помощью регрессионной модели.

При выборе оптимальной МЛР важную роль играет сумма квадратов остатков (СКО)

 ISSN 1562-9945

$$\sum_{i=1}^{n} \hat{U}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \qquad (6)$$

Интуитивно ясно, что чем лучше МЛР описывает исследуемый процесс, тем меньше должна быть СКО.

Под выбором оптимальной модели мы понимаем следующую задачу.

Предположим, что имеется ***m*** переменных ***X***, которые, возможно, влияют на переменную ***Y***.

Требуется из этого множества выбрать подмножество, содержащее ***k*** переменных, которые наилучшим образом объясняют наблюдаемые значения зависимой переменной ***Y***.

Сокращение числа независимых переменных позволяет не только уменьшить размерность модели, но и избежать эффектов связанных, например, с мультиколинеарностью. Поэтому ухудшение качества используемой модели происходит как при исключении из неё существенной независимой переменной, так и при включении избыточных переменных. [4]

Были проанализированы 5 различных методов для определения оптимальной МЛР:

- метод всех возможных регрессий с исправленным коэффициентом детерминации в качестве критерия оптимальности (MR2);[1,4]

- метод всех возможных регрессий с исправленным коэффициентом детерминации в качестве критерия оптимальности и оценкой значимости коэффициентов МЛР (на основе t-статистики) (MR2t);

- метод всех возможных регрессий с использованием статистики Маллоуза в качестве критерия оптимальности (Mlz);[3,4]

- метод последовательного исключения (BWE) [2,4];

- пошаговый метод (SWP) [2,4] .

Метод всех возможных регрессий.Перебираются все возможные модели первого порядка, и определяется наилучшая модель среди них, затем перебираются все возможные модели второго порядка и выбирается наилучшая из них и так далее, вплоть до МЛР порядка *m*. Определяется наилучшая модель среди всех возможных. Этот метод требует построения каждого из всех возможных регрессионных

уравнений. Если общее количество независимых переменных m, то, поскольку для каждой переменной X есть всего две возможности: либо входить, либо не входить в уравнение, всего нужно проанализировать 2^m уравнений МЛР. Для выбора оптимальной МЛР применялись:

1) Критерий минимума исправленного коэффициента детерминации.

Исправленный коэффициент детерминации рассчитывается следующим образом [1]:

$$\tilde{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k}, \qquad (7)$$

где $\boldsymbol{n}$ - объём выборки, $\boldsymbol{k}$ - число независимых переменных, включённых в МЛР,

а $\boldsymbol{R^2}$ обозначает коэффициент детерминации [1]:

$$R^2 = \frac{\hat{\beta}' X' Y - n\overline{Y}^2}{Y'Y - n\overline{Y}^2}. \qquad (8)$$

В последней формуле:

$\hat{\beta}$ - $[(k+1)\times 1]$ вектор рассчитанных оценок параметров МЛР,

$\boldsymbol{X}$ - $[(n\times m]$ матрица наблюдаемых значений независимых переменных,

$\boldsymbol{Y}$ - $[n\times 1]$ вектор зависимых переменных (чёрточка сверху обозначает, как обычно, среднее арифметическое, а штрих – операцию транспонирования).

2) Нами исследовался также алгоритм, в котором критерий минимума исправленного коэффициента детерминации был дополнен проверкой значимости оценок коэффициентов регрессии модели, выбранной в качестве оптимальной. Если некоторые из оценок коэффициентов оказываются незначимыми, то выбранная модель усекается путем отбрасывания членов с незначимыми коэффициентами. Значимость оценок проверялась с помощью $\boldsymbol{t}$ – статистики

$$t = \frac{\hat{\beta}_l}{\hat{\sigma}_{\hat{\beta}_l}},\ l = 1, ..., k \qquad (9)$$

где $\hat{\sigma}_{\hat{\beta}_l}$ - обозначение выборочной дисперсии оценок $\hat{\beta}_l$, $l = 1, 2, ..., k$.

 ISSN 1562-9945

Так как статистика ***t*** имеет распределение Стьюдента в случае, когда ошибки МЛР распределены нормально, то оценка коэффициента $\hat{\beta}_l$ не считается значимой, если

$$\left|t_j\right| \le t_{\alpha/2}, \qquad (10)$$

где - $t_{\alpha/2}$ критическая точка распределения Стьюдента для заданного уровня значимости α.

3) Критерий минимума статистики Маллоуза.

Статистика Маллоуза C_p вычисляется по формуле [2,4]:

$$C_p = ns_k^2 / s_m^2 - (n - 2(k+1)), \qquad (11)$$

где s_k^2 - СКО полученная для модели, содержащей ***k*** регрессоров и s_m^2 - СКО, полученная для модели, включающей все возможные ***m*** регрессоров.

Метод последовательного исключения (BACKWARD ELIMINATION). Метод последовательного исключения более экономичный, чем метод всех регрессий, поскольку в нем делается попытка исследовать только лучшие регрессионные уравнения, содержащие определенное число переменных. Основные шаги этого метода сводятся к следующему:

1. Рассчитывается регрессионное уравнение, включающее все переменные.

2. Вычисляется величина частного F-критерия [1]:

$$F(1, v, 1-\alpha) = \left\{ t(v, 1-\frac{\alpha}{2}) \right\}^2, \qquad (12)$$

имеющего распределение Фишера. Для каждой предикторной переменной, в предположении, так как будто она была последней переменной, введенной в регрессионное уравнение. (В формуле (12) ***t*** - значение ***t***-статистики, ***v*** - число степеней свободы, α - уровень значимости.)

3. Наименьшая величина частного ***F***-критерия F_j, сравнивается с заранее выбранным критическим значением, например F_0.

а) Если $F_j < F_0$, то переменная X_j, которая обеспечила достижение уровня F_j, исключается из рассмотрения и производится перерасчет уравнения регрессии с учетом переменных, оставшиеся, переходим к пункту а).

б) Если $F_j > F_0$, то регрессионное уравнение оставляют таким, как оно было рассчитано.

Пошаговый метод (STEPWISE PROCEDURE).Этот метод также относится к быстрым методам. Вначале в оптимальную модель не входит ни одна независимая переменная.

1. Находим, какая из независимых переменных Xj, $j=1, 2,..,m$ сильнее всего коррелирует с переменной Y. Добавляем эту независимую переменную в оптимальную модель.

2. Проверяем значимость коэффициентов независимых переменных, вошедших в модель. Удаляем переменные, имеющие незначимые коэффициенты.

3. Если остались еще нерассмотренные независимые переменные, переходим к пункту 1.

Исследование эффективности различных методов выбора оптимальной модели МЛР.Данное исследование проводилось с помощью имитационного моделирования.

На ЭВМ моделировалось m случайных последовательностей, которые использовались в качестве значений независимых переменных. Из них выбирались k последовательностей в качестве значений переменных участвующих в построении МЛР (1). Затем предполагалось, что неизвестно какие и сколько из m независимых переменных входят в истинную модель. Таким образом, поиск модели осуществлялся с использованием всех возможных независимых переменных с применением каждого из 5 описанных выше критериев оптимальности. Для выяснения качества каждого из рассмотренных методов оптимизации проводились численные эксперименты, в которых поиск оптимальной модели производился многократно по статистически независимым данным. А именно, генерировались N (число экспериментов) наборов из m последовательностей, каждая из которых содержала по n (объём выборки) данных. В каждом из N экспериментов определялась наилучшая модель на основе перечисленных выше пяти критериев. Так как истинная МЛР заранее известна, описанная методика позволяет оценить процент правильных идентификаций каждым из исследуемых методов. Эта оценка будет достаточно надёжной, если N достаточно велико. Для проведения численных экспериментов нами было разработано в среде МАТЛАБ соответствующее программное обеспечение.

 ISSN 1562-9945

В таблице 1 приведены результаты описанных выше численных экспериментов, полученные при $N = 1000$ и $n = 100$.

Приведенные данные показывают, что среди методов, основанных на переборе всех моделей, наилучшие результаты даёт процедура, основанная на критерии Маллоуза. Этот критерий для всех рассмотренных случаев даёт 100% правильных идентификаций истинной модели.

Широко используемый в практике регрессионного анализа исправленный коэффициент детерминации, даже дополненный проверкой значимости оценок коэффициентов модели, дает удовлетворительные результаты только при небольшом количестве возможных регрессоров (порядка 5-7). Общим недостатком этой группы методов является быстрое увеличение количества возможных моделей при увеличении числа возможных регрессоров m. (напомним, что количество всех возможных МЛР составляет 2^m). Поэтому при $m > 15$ перебор всех возможных регрессий становится невозможным.

На основе данных таблицы № 1 построены графики, показанные на рисунке 1.

Таблица 1

Процент совпадения правильных идентификаций моделей для разных методов поиска МЛР в зависимости от параметров m, при значении N = 1000 и при использовании двоих процессорных ядер

k	*m*	Метод поиска оптимальной модели				
		MR2	***MR2t***	***Mlz***	***BWI***	***SWP***
2	3	67.6	97.9	100	98.8	97.9
3	4	68.3	97.4	100	99	97.40
5	5	100	100	100	100	99.7
5	6	68.1	97.7	100	98.6	97.7
5	7	45.9	87	100	98.1	96.5
5	8	31.3	71.6	100	96.7	94.2
5	9	20.6	58.4	100	95.5	93
5	10	13.5	43.7	100	94.3	91.4
5	11	8.30	31.9	100	93	89.9
5	12	6.5	27.1	100	91.7	88.4
5	13	4.50	18.9	100	90.7	86.7
5	14	3	14.2	100	89.8	85.7
5	15	-	-	-	88.4	83
5	16	-	-	-	87.2	81.6

5	20	-	-	-	83.2	77.4
5	30	-	-	-	71.4	67.7
5	40	-	-	-	61.7	58.8
5	50	-	-	-	52.4	52.4
5	70	-	-	-	38.2	40.7

Широко используемый в практике регрессионного анализа исправленный коэффициент детерминации, даже дополненный проверкой значимости оценок коэффициентов модели, дает удовлетворительные результаты только при небольшом количестве возможных регрессоров (порядка 5-7). Общим недостатком этой группы методов является быстрое увеличение количества возможных моделей при увеличении числа возможных регрессоров m. (напомним, что количество всех возможных МЛР составляет 2^m). Поэтому при $m > 15$ перебор всех возможных регрессий становится невозможным.

Также было проведено исследование быстродействия изучаемых алгоритмов идентификации оптимальной МЛР.

Разработанный программный продукт с целью ускорения вычислений программно распределяет задачи на все физические процессорные ядра, то есть создает столько независимых параллельно работающих потоков выполнения, сколько процессорных ядер на компьютере. Разработанное программное приложение прошло тестирование на ЭВМ с двумя процессорными ядрами (тестирование проводилось на компьютере с параметрами: RAM 3,25 GB, процессор Intel (R) Core (TM) 2 Duo CPU E8400@3.00 GHz, 3.00 GHz).

Данные, приведенные в таблице №2, показывают, что с помощью программного распараллеливания процессорных вычислений было достигнуто ускорение проведения расчетов приблизительно в два раза при увеличении количества процессорных ядер, задействованных в вычислениях с одного до двух.

Рисунок 2 иллюстрирует данные, приведенные в таблице № 2.

Методы всех возможных регрессий показывают низкое быстродействие, что связано с тем, что для нахождения оптимальной модели этим методам необходимо построить 2^m моделей и для каждой применить критерий оптимальности.

 ISSN 1562-9945

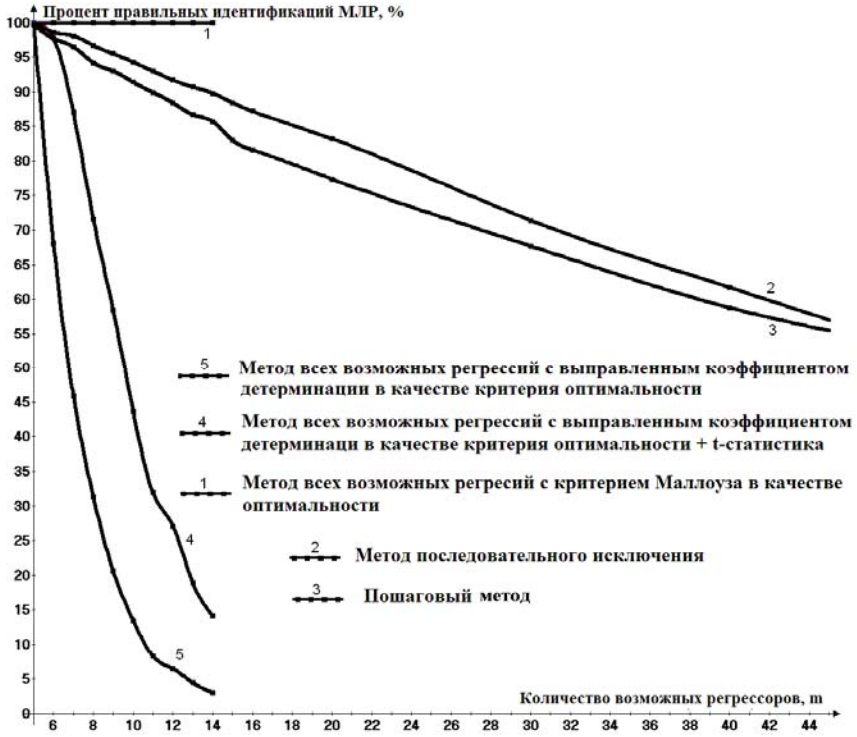

Рисунок 1 - Процент правильных идентификаций МЛР для различных методов, при фиксированном параметре k = 5

Таблица 2

Время выполнения вычислений (в секундах) для разных методов поиска МЛР в зависимости от параметра ***m*** при значении ***n*** =100 и ***N*** = 1000

k	***m***	Количество процессорых ядер, задействованных в вычислениях									
		1					2				
		Метод проведения вычислений									
		MR2	***MR2t***	***Mlz***	***BWI***	***SWP***	***MR2***	***MR2t***	***Mlz***	***BWI***	***SWP***
2	3	1,44	1,65	0,72	0,77	2,40	0,72	0,86	0,38	0,42	1,15
3	4	2,28	2,89	1,46	0,78	2,95	1,21	1,46	0,76	0,46	1,64
5	5	4,15	4,7	2,26	0,64	4,28	2,46	2,42	1,12	0,4	2,35
5	6	7,81	9,17	3,59	0,9	5,47	4,174	4,45	1,79	0,5	2,93
5	7	15,29	15,93	6,37	1,11	6,82	8,023	8,36	3,2	0,64	3,63
5	8	29,73	30,75	12,0	1,37	8,31	15,76	15,99	6,03	0,78	4,54
5	9	59,6	61,14	23,02	1,68	9,78	31,29	30,8	11,92	0,9	5,21
5	10	119,45	122,1	45,04	1,99	11,6	62,14	61,38	23,75	1,07	6,14
5	11	242,6	287,6	90,24	2,31	13,4	122,5	127,3	45,1	1,22	7,31
5	12	490,65	493,72	186	2,65	15,5	247,4	250,3	93,2	1,39	8,18
5	13	1023,9	1002	390	3,0	17,9	509,3	512,4	186,1	1,57	9,55
5	14	2052,5	2055	787	3,39	19,9	1022,6	1027	390,1	1,76	10,64

5	15	-	-	-	3,83	22,3	-	-	-	1,97	11,62
5	16	-	-	-	4,27	24,6	-	-	-	2,19	12,81
5	20	-	-	-	6,14	36,4	-	-	-	3,35	17,6
5	30	-	-	-	13,3	70,3	-	-	-	6,75	35
5	40	-	-	-	26,4	117	-	-	-	13,15	58,4
5	50	-	-	-	40,7	191	-	-	-	17,31	85,8
5	70	-	-	-	64,3	360	-	-	-	33,73	154,6

Приведенные данные (таблица №2) показывают, что метод последовательного исключения дает наилучшие показатели быстродействия. Этот метод может быть использован для поиска оптимальной МЛР, когда количество возможных регрессоров велико ($m > 20$).

Пошаговый метод менее быстродействующий, чем метод последовательного исключения, но все же он значительно превосходит по быстродействию методы всех возможных регрессий.

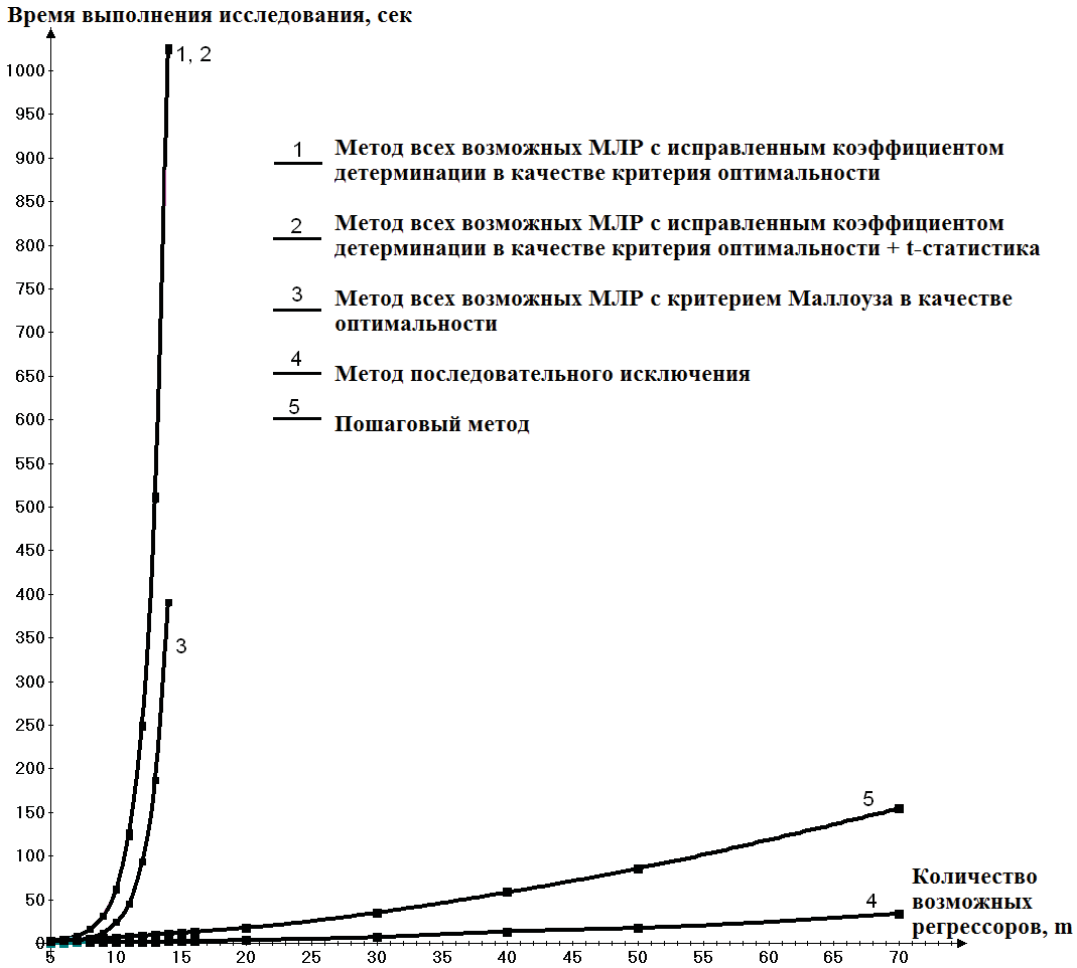

Рисунок 2 - Время выполнения вычислений разными методами в зависимости от количества возможных регрессоров m, при фиксированном параметре k = 5, при использовании двух ядер процессора

 ISSN 1562-9945

Выводы. Исследование представленных методов выбора оптимальной модели множественной линейной регрессии было проведено с помощью программного обеспечения разработанного в среде MATLAB.

Анализ проводился методом имитационного моделирования.

В случае, когда количество возможных независимых переменных меньше 15, метод всех возможных регрессий со статистикой Маллоуза в качестве критерия оптимальности наиболее часто выбирает в оптимальную модель те и только те независимые переменные, которые входят в истинную модель. Но методы всех возможных регрессий неприменимы в случае, если количество возможных независимых переменных более 14. В этом случае необходимо использовать быстрые методы: метод последовательного исключения и пошаговый. Оба этих метода обладают высоким быстродействием и точностью.

ЛИТЕРАТУРА

1. Дрейпер Н., Смит Г. – Прикладной регрессионный анализ: В 2-х кн. 2/Пер. с англ. – 2-е изд., перераб. и доп. – М.: Финансы и статистика. 1987. – 351с.
2. Green William H. Econometric Analysis. Upper Saddle River, NJ: Prentice Hall, 2003.
3. Jonston J. and John DiNardo. Econometric Methods, 4th td. New York: McGraw Hill, 1997.
4. Rawlings J.O., S.G.Pantula, D.A.Dickey. Applied Regression Analysis. A Research Tool.Second edition – New York: Springer, 2001. - 671 p. - ISBN 0387984542