МЕТОДИЧЕСКИЕ ПОДХОДЫ К ИССЛЕДОВАНИЮ МНОГОМЕРНЫХ ЗАВИСИМОСТЕЙ НА ПРИМЕРЕ ФОНДОВОГО РЫНКА

УДК 311.2

Марина Анатольевна Скорик,
к.э.н., доцент, доцент каф. Математической статистики и
эконометрики Московского государственного университета
экономики, статистики и информатики (МЭСИ)
Тел.: (495) 442-71-77
Эл. почта: maskorik@mesi.ru

Андрей Геннадиевич Нефедов,
старший преподаватель каф. Математической статистики и
эконометрики Московского государственного университета
экономики, статистики и информатики (МЭСИ)
Тел.: (495) 442-71-77
Эл. почта: anefedov@mesi.ru

Многопризнаковая природа реальных явлений и процессов объясняет постоянно возрастающий интерес к методам многомерного статистического анализа, которые по праву являются интеллектуальным инструментарием современного исследователя. В рамках годового курса «Многомерные статистические методы» традиционно рассматриваются такие основные разделы как «Множественный корреляционно-регрессионный анализ», «Методы снижения размерности», «Методы многомерной классификации». Теоретический материал сопровождается практическими занятиями в компьютерном классе. Широкие возможности для выполнения предусмотренных курсом лабораторных работ предоставляет программа «SEMESTR» для пакета Mathcad, предназначенная для автоматизации обработки данных на основе методов многомерного анализа. В статье излагаются методические основы исследования многомерных взаимосвязей на примере фондовых индексов.

Ключевые слова: многомерный анализ, регрессионный анализ, компонентный анализ, фондовые индексы, размерность признакового пространства, Mathcad.

Marina A. Skorik,
Candidate of Economic Sciences, Associate Professor, Department of Mathematical Statistics and Econometrics, Moscow State University of Economics, Statistics and Informatics (MESI)
Tel.: (495) 442-71-77
E-mail: maskorik@mesi.ru

Andrey G. Nefedov,
Senior Lecturer, the Department of Mathematical Statistics and Econometrics, Moscow State University of Economics, Statistics and Informatics (MESI)
Tel.: (495) 442-71-77
E-mail: anefedov@mesi.ru

APPROACH TO THE RESEARCH OF THE MULTIVARIATE RELATIONSHIP BY THE EXAMPLE OF RUSSIAN STOCK MARKET

Many-dimensional nature of real occurrences and processes explains the actually increasing interest to the multivariate statistical analysis as a sophisticated toolset of the advanced researcher. Within the framework of the annual course «Multivariate statistical methods» such subsections as «Multivariate regression analysis», «Methods of the reduction of the variable space dimensionality», «Multivariate classification» are taught. All theoretical foundations are supported by praxis in computer class. The Mathcad based program «SEMESTR» provides the wide range of possibilities for performing laboratory works by course due to automation of the multivariate data processing. The article is devoted to the methodical principles of the study of multivariate relations between stock indexes.

Keywords: regression analysis, factor analysis, multivariate analysis, stock indexes, dimensionality of factors space, Mathcad.

1. Введение

Итоги анализа исходных данных, с применением современных статистических пакетов, представляют собой окончательные результаты машинной обработки, без промежуточных расчетов, которые остаются «как бы закрытыми» для студентов, выполняющих семестровые задания по курсу «Многомерные статистические методы». В то же время для эффективности учебного процесса ознакомление именно с этими промежуточными расчетами столь же необходимо (и полезно), как и получение окончательных результатов и их интерпретация.

Проблема в том, что уже при трех объясняющих признаках трудоемкость вычислений, производимых вручную, существенно повышается. Именно этим была продиктована необходимость создания программы на встроенном в Mathcad языке программирования и внедрения ее в учебный процесс. Программа «SEMESTR» предназначена для автоматизации обработки данных методами корреляционно-регрессионного и компонентного анализа.

Авторами был проанализирован широкий спектр исходных данных и выбраны такие подмножества, на примере которых удалось удачно проиллюстрировать как работу методов, так и самой программы, а также принципы содержательной интерпретации полученных результатов.

Следует сразу подчеркнуть, что данная программа не предназначена для решения задач большой размерности. Авторы не стремились создать конкурентоспособный программный продукт. Преследовалась сугубо учебная цель: проиллюстрировать (с помощью промежуточных результатов) работу некоторых методов многомерного статистического анализа, а также возникающие при этом проблемы и пути их решения.

2. Методика исследования зависимостей в курсе MCM

Экономико-статистический анализ экономических процессов и явлений включает в себя комплексное изучение зависимостей между рассматриваемыми показателями, что предполагает необходимость использования многомерных статистических методов. Задача определения формы зависимости одного признака от группы других приводит к необходимости построения уравнения регрессии. При этом неявно предполагается, что исследуемые признаки имеют совместное многомерное нормальное распределение. Такое предположение предоставляет исследователю ряд преимуществ. В случае многомерного нормального распределения отдельные признаки либо независимы между собой, либо зависимость (если она существует) является линейной, а это значит, что понятия некоррелированности и независимости совпадают.

Тогда можно ограничиться рассмотрением только линейных моделей, для которых разработан эффективный математический аппарат и создано соответствующее программное

обеспечение. Кроме того, упрощается процесс формализации модели, обоснование выбора типа модели, интерпретация результатов и их использование для выработки рекомендаций. Матрица парных коэффициентов корреляции достаточно хорошо характеризует тесноту связи между признаками. Тем самым создаются предпосылки для эффективного использования методов многомерного статистического анализа, опирающихся на матрицу парных коэффициентов корреляции, а именно корреляционного, регрессионного и компонентного анализа. [1]

Работа начинается с экономической постановки задачи, сбора данных, анализа имеющейся информации, содержательного исследования взаимосвязей и зависимостей между рассматриваемыми показателями. При этом уточняется набор признаков, участвующих в модели и формируются рабочие гипотезы, которые в дальнейшем будут проверяться количественными расчетами с использованием методов многомерного статистического анализа.

3. Корреляция как метод разведочного анализа данных

Предполагается следующая последовательность действий. Сначала строится корреляционная матрица, которая позволяет получить предварительное представление о характере исследуемых взаимосвязей между показателями (теснота и направление). Оценить значимость можно как по самим значениям коэффициентов корреляции, так и по соответствующим значениям *t*-статистики. Для оценки дублирования информации можно построить матрицу частных коэффициентов корреляции, указав номер зафиксированной переменной.

Из-за наличия статистической связи между факторными признаками они оказывают влияние на результирующий признак как непосредственно, так и косвенно, через другие факторные признаки. Становится невозможным однозначно охарактеризовать его изменение, так как между факторами идет борьба за информационную нагрузку в объяснении результирующего призна-

ка. Ситуация особенно обостряется, если между несколькими факторными признаками существует почти функциональная зависимость, т.е. один из признаков является линейной комбинацией других. В этом случае следует учитывать возможность появления мультиколлинеарности (см. табл.1), причем отсутствие явных признаков этого (коэффициентов корреляции между потенциальными регрессорами, по модулю больших, чем 0,8), не гарантирует отсутствия самого явления.

Наличие мультиколлинеарности приводит к резкому ухудшению качества уравнения регрессии, его статистических характеристик: формальные результаты входят в противоречие с выводами содержательного анализа. Кроме того, поскольку факторные признаки сильно конкурируют за информационную нагрузку для объяснения результирующего признака, то уравнение становится неустойчивым, т.е. незначительное изменение состава выборки может привести к сильному изменению отдельных коэффициентов уравнения. А ухудшение отношения между объемом выборки и количеством переменных, влияющее на значимость самого уравнения, ставит под сомнение адекватность его исследуемому реальному процессу. [6]

Все вышеизложенное не позволяет использовать полученное уравнение, поэтому от мультиколлинеарности следует избавиться до начала его построения. Набор показателей проверяется на наличие линейно связанных между собой пар признаков, указывается порядок включения признаков в уравнение регрессии на следующем этапе. Тем самым уточняется перечень регрессионных моделей, которые предстоит построить, отказавшись от заведомо бесперспективных. Это позволяет уменьшить объем вычислений и затраты труда на анализ и интерпретацию результатов.

4. Выявление взаимосвязей на основе регрессионного анализа

Следует иметь в виду, что составление наиболее информативного набора регрессоров для объяснения вариации *Y* не эквивалентно выбору регрессоров, наиболее тесно связанных с *Y* в отдельности. Необходимо

Таблица 1

Парные коэффициенты корреляции мировых фондовых индексов в 2005–14 г.

Индекс	RTSI	BOVESPA	CAC 40	DAX 30	FTSE 100	Hang Seng	IBEX 35	D&J	IPC	BSE	XU 100	JSE	PX 50
RTSI	1,00												
BOVESPA	0,65	1,00											
CAC 40	0,52	-0,17	1,00										
DAX 30	0,64	0,50	0,31	1,00									
FTSE 100	0,69	0,29	0,60	0,87	1,00								
Hang Seng	0,81	0,82	0,23	0,80	0,66	1,00							
IBEX 35	0,58	0,06	0,88	0,18	0,38	0,34	1,00						
D&J	0,52	0,27	0,33	0,92	0,91	0,62	0,08	1,00					
IPC	0,49	0,76	-0,25	0,79	0,54	0,74	-0,25	0,68	1,00				
BSE	0,60	0,85	-0,12	0,80	0,56	0,87	-0,04	0,62	0,93	1,00			
XU100	0,42	0,64	-0,17	0,79	0,62	0,69	-0,27	0,71	0,92	0,88	1,00		
JSE	0,46	0,64	-0,15	0,88	0,63	0,73	-0,21	0,79	0,95	0,89	0,89	1,00	
PX50	0,61	-0,02	0,93	0,18	0,47	0,29	0,92	0,16	-0,27	-0,11	-0,22	-0,24	1,00

учитывать взаимосвязь регрессоров, поскольку, чем теснее эти связи, тем менее информативен их набор. Например, при наличии функциональной связи между x_1 и x_2 информативность набора (x_1, x_2) равна информативности каждого – x_1 и x_2 – в отдельности. Поэтому наборы информативных признаков удобно определять с помощью пошаговых процедур. [2]

При этом проще всего организуется схема исключения. Сначала строится МНК-уравнение на все объясняющие признаки. Далее проверяется значимость уравнения и отдельных коэффициентов регрессии (как правило, берется $\alpha = 0,05$, но исследователь вправе изменить уровень значимости исходя из содержательного смысла задачи). При наличии незначимых регрессоров ($|t_{ij}| < t_{кр}$) необходимо выбрать наименее значимый ($\min |t_{ij}|$) и исключить его из уравнения. На следующем шаге МНК-уравнение строится на оставшиеся признаки. Процесс повторяется до тех пор, пока в уравнении не останутся только значимые регрессоры.

Например, построим уравнение зависимости российского индекса РТС от фондовых индексов развитых и развивающихся стран: США (D&J), Великобритании (FTSE100), Германии (DAX30), Франции (CAC40), Испании (IBEX35), Чешской республики (PX50), Турции (XU100), Мексики (IPC), Бразилии (BOVESPA), Индии (BSE), Китая (HangSeng), ЮАР (JSE) за 2005-2014 гг.:

$$RTSI = -1570,13 + 0,01 \cdot BOVESPA - 0,07 \cdot CAC40 - 0,09 \cdot DAX30 + 0,39 \cdot FTSE100 + t_{i:} \begin{matrix} (-45,87) & (12,61) & (-2,72) & (-7,34) & (16,59) \\ +0,02 \cdot HangSeng - 0,06 \cdot IBEX35 - 0,04 \cdot D\&J + 0,01 \cdot IPC - 0,02 \cdot XU100 + 0,03 \cdot BSE + 1,21 \cdot PX50 + 0,02 \cdot JSE \end{matrix} \quad (1)$$

$$R^2 = 0,8994, S^2 = 138,62$$

При использовании схемы последовательного включения можно реализовать более квалифицированный подход. По матрице R по строке (столбцу), соответствующей результативному признаку, выбирается наиболее коррелированный с Y -ом регрессор и строится МНК-уравнение на него, проверяется значимость. Далее исследователь возвращается в корреляционный анализ и по матрице частных коэффициентов корреляции по строке, соответствующей результативному признаку, выбирает наиболее коррелированный показатель (таким образом, обеспечивается минимальное дублирование информации с ранее включенными в уравнение признаками). Этот регрессор вводится в модель. Процесс прекращается, если введен незначимый регрессор.

Комбинируя два изложенных подхода можно организовать схему «включение-исключение». Первый шаг в этом случае проводится по схеме включения, а на каждом следующем сначала проверяется, можно ли включить какой-то претендент в модель, а затем проверяется, можно ли исключить незначимый регрессор из модели. Процесс прекращается, если ни один признак из не включенных нельзя включить в уравнение и ни один из ранее включенных нельзя вывести из уравнения.

Коррелированность исходных признаков приводит к избыточной информации при включении в модель большого числа регрессоров. Решение становится неус-

тойчивым, незначительное изменение состава выборки (включение новых объектов, различающихся значениями признаков) может вызвать кардинальное изменение модели. Наряду с этим недостаточная информативность регрессоров приводит к незначимости уравнения (и отдельных его параметров), поэтому повышать информативность всего набора регрессоров за счет включения в него малоинформативных регрессоров можно только после устранения всех остальных причин слабой связи. Если существенные признаки не включены в модель, то коэффициенты регрессии получаются смещенными. [1]

На практике часто используют несколько схем для построения уравнения регрессии одного и того же результативного признака на данный набор объясняющих факторов. Несовпадение результатов различных схем также указывает на потерю устойчивости уравнения регрессии из-за мультиколлинеарности и, как следствие, на нецелесообразность включения в уравнение дополнительных регрессоров. Таким образом, сравнение результатов, полученных по разным схемам, позволяет исследователю определить наиболее целесообразное уравнение регрессии. При этом последовательно анализируется каждый шаг построения уравнения регрессии; проверяется значимость как самого уравнения, так и отдельных коэффициентов.

Таблица остатков регрессионного анализа позволяет дать характеристику объектам с точки зрения решаемой задачи, условно разделив их на «передовые» (у которых абсолютные и относительные отклонения в лучшую сторону, т.е. существенно выше нуля), «отстающие» (отклонения в худшую сторону, т.е. ниже нуля) и «средние», занимающие промежуточное положение (отклонения близки к нулю). При этом последовательно анализируется каждый шаг построения уравнения регрессии. Если построенное уравнение значимо, проверяем значимость отдельных коэффициентов. Указываются причины и характер изменения значимости коэффициентов регрессии при ранее введенных в уравнение признаках. Естественно, проверке подлежат все коэффициенты, участвовавшие в проверке значимости уравнения.

Далее можно построить ковариационную матрицу $V(b)$ для коэффициентов уравнения регрессии: выборочные коэффициенты – случайные величины, и поскольку признаки X_j коррелируют между собой, то коэффициенты регрессии реагируют на эту взаимосвязь, коррелируя, в свою очередь, друг с другом. В случае, когда уравнение признано хорошим, адекватным исследуемому процессу, можно переходить к построению доверительных интервалов для коэффициентов регрессии и для результативного признака. На этом формальный этап решения задачи заканчивается, и переходят к содержательной интерпретации результатов.

При проведении интерпретации оценивается не только содержательный смысл модели, но и информативность, например, с помощью множественного коэффициента корреляции (детерминации) этого окончательного уравнения по сравнению с аналогичным, построенным по полному набору исходных объясняющих показателей. Потери информации (ΔR^2) могут быть достаточно

большими и тогда целесообразно перейти к регрессии на главные компоненты. [3]

Необходимо при этом учитывать, что пошаговая регрессия производит обработку признаков-регрессоров, не учитывая экономического содержания признаков и самой задачи. Поэтому, если исследуемое уравнение, по мнению исследователя, не соответствует экономическому смыслу, целесообразно повторить расчет по программе множественного регрессионного анализа с другим (новым), скорректированным набором признаков.

В ходе исследования возможна ситуация, когда построенное уравнение, являясь безупречным с формальных позиций, не устраивает пользователя по содержательным соображениям. Например, он считает, что на результативный признак должны оказывать влияние не только те объясняющие признаки, которые вошли в окончательное уравнение регрессии, но и еще некоторые, влияние которых в построенных моделях статистически не подтверждается.

В принципе понятно, почему это произошло. Объясняющие признаки только предполагаются независимыми. В действительности они коррелируют между собой, т.е. дублируют друг друга. Поэтому, когда они вместе оказываются в уравнении регрессии, они начинают «конкурировать» между собой «за право» объяснять результативный признак. В результате этого самые слабые (наименее значимые) регрессоры покидают уравнение. На этих идеях и основаны пошаговые процедуры. Другого результата исследователь получить в рамках классического регрессионного анализа не может. Он вынужден использовать более тонкую методику исследования. Такая возможность представляется ему при переходе от набора исходных признаков к новому набору «агрегированных» признаков – главных компонент.

5. Поиск латентных факторов на основе компонентного анализа

Компонентный анализ используется в прикладных исследованиях с несколькими *целями*:

- определяются некоторые объективно существующие закономерности, непосредственно не наблюдаемые, которые могут быть содержательно интерпретированы, и тем самым позволяют исследователю лучше представить процесс и скорректировать дальнейший план исследования.

- в результате применения компонентного анализа исследователь получает матрицу нагрузок для анализа взаимосвязи признаков и найденных главных компонент, и на ее основе – индивидуальные значения главных компонент на объектах, которые используются при построении диаграммы рассеяния, как для решения задачи классификации объектов в ортогональном пространстве главных компонент (вместо косоугольного пространства исходных признаков), так и для сравнения с результатами ранее проведенного исследования выборки на однородность, проверки гипотезы о нормальном распределении.

- при затруднении с содержательной интерпретацией полученных главных компонент, они могут быть ис-

пользованы как начальные приближения для решения задачи факторного анализа.

- главные компоненты можно использовать для построения на них уравнения регрессии результативного признака (с дальнейшим пересчетом на исходные признаки). Поскольку использование всех главных компонент приводит к уравнению, тождественному с уравнением, построенным на все исходные признаки, на практике, как правило, используют лишь несколько первых наиболее весомых главных компонент. При этом наиболее существенная информация о процессе сохраняется, а случайная отбрасывается.

- главные компоненты можно применять для восстановления пропущенных значений исходных признаков. Эта задача недостаточно обоснована теоретически, но на практике иногда дает неплохие результаты. Восстановленные исходные значения можно сравнить с фактическими значениями X_{ij} и проанализировать характер отклонений (сравнивая эти «гипотетические» значения исходных признаков с фактическими), охарактеризовав положение дел на каждом объекте. Данный подход может быть использован для исследования «узких мест» на предприятиях и полезен при выработке рекомендаций по их устранению. [3]

Необходимо дать объяснение причинам полученных результатов предварительной классификации объектов на основе диаграммы рассеяния. Если четко проследивается наличие двух, трех или большего количества облаков, то целесообразно разбить всю выборку на несколько (по числу облаков), включая в каждую из выборок лишь определенные объекты. В этом случае представляет интерес сравнение результатов, полученных по каждой группе в отдельности, с результатами, полученными по всему множеству, что укажет специфику каждой группы.

При анализе главных компонент особое внимание следует уделить вопросу их содержательной интерпретации. Хотя найденные компоненты и являются абстракциями, они отражают объективно существующие, но непосредственно не наблюдаемые закономерности, присущие исследуемому процессу. Программа рассчитывает и выдает исследователю собственные вектора и собственные числа: для матрицы ковариаций, если нормировка исходных признаков не производится; для матрицы корреляции, если затребована операция центрирования и нормирования. Причем результаты различаются из-за проводимого линейного преобразования. Следует иметь в виду, что качественная картина лучше просматривается для стандартизованных переменных, а количественная (например, при построении уравнения регрессии на главные компоненты) – для нестандартизованных.

Представляют интерес вопросы: с какими именно признаками связана та или иная главная компонента и почему именно эти линейные комбинации исходных признаков в наибольшей мере влияют на взаимосвязь (корреляцию) между исходными признаками? При этом, используя свойство ортогональности матрицы, составленной из собственных векторов, можно рас-

смотреть расположение исходных признаков в пространстве главных компонент и оценить «расстояние» между признаками.

Если таких компонент наберется три, то можно в целях наглядности построить три графика в координатах (f_1, f_2) , (f_1, f_3) , (f_2, f_3) . Такие графики помогут выявить группы близких между собой исходных признаков в координатах данной пары компонент. При этом необходимо следить за соответствием содержательного смысла и формальных результатов, а в случае расхождения указать причину, по которой одна и та же пара признаков может быть близкой с точки зрения одной пары главных компонент и далекой для другой пары главных компонент. При неудовлетворительной, с точки зрения исследователя, интерпретации главных компонент, расчет целесообразно повторить со скорректированным набором признаков.

Главные компоненты отражают устойчивые закономерности, а следовательно, менее чувствительны к изменению состава выборки. Поэтому можно рассчитать значения всех главных компонент на каждом объекте, а затем, используя матрицу $U^{-1} = U^T$, по значениям главных компонент (и известным значениям имеющихся признаков) получить значения пропущенных признаков. Найденные таким образом значения признаков, в среднем, ближе к реальным, чем, например, полученные в результате усреднения или экспертного оценивания. [4]

Используя эти новые признаки в качестве объясняющих, можно с помощью обычных процедур регрессионного анализа построить более приемлемое уравнение регрессии. Идея этого метода состоит в том, что рассчитываются значения главных компонент на каждом объекте, а затем к этим новым признакам применяется один из методов регрессионного анализа: матричный алгоритм или пошаговая схема. Если пошаговые процедуры предлагают исследователю несколько приемлемых уравнений регрессии, то окончательный выбор он осуществляет, опираясь, как на формальные результаты (значимое уравнение и значимые коэффициенты при регрессорах, максимальный коэффициент детерминации, минимальная ошибка аппроксимации, иногда минимальный по модулю свободный член уравнения – вклад неучтенных факторов в значение результативного признака, таблица остатков и т.д.), так и на содержательные.

Свободный коэффициент полученного уравнения в общем случае равен среднему значению результативного признака. В том случае, если результативный признак стандартизован, то свободный коэффициент практически равен нулю. Это означает, что плоскость регрессии проходит через центр выборки, в который помещено начало новой системы координат, построенной по главным компонентам. Можно упростить уравнение, удалив из него незначимые составляющие. При этом следует помнить о том, что свойство ортогональности (некоррелированности) главных компонент позволяет исключить какую-либо из них из уравнения без пересчета оставшихся коэффициентов регрессии. При этом

необходимо следить за соответствием содержательного смысла результативного признака смыслу включенных в уравнение главных компонент – обобщенных показателей.

Уравнение регрессии можно строить на все главные компоненты или на несколько наиболее весомых. Дело в том, что наиболее коррелированной с результативным признаком может быть не первая или вторая главные компоненты, т.е. не обязательно самые весомые. Но с другой стороны, именно первые, наиболее весомые главные компоненты, наиболее информативны для исследуемого процесса, поэтому пренебречь ими нельзя, даже если они слабо коррелированы с результативным признаком. Если возникла такая ситуация, это означает, что неверно выбран состав факторов для объяснения данного результативного признака.

На практике результаты получаются, как правило, хорошие, т.е. несколько первых, наиболее весомых, главных компонент являются и наиболее коррелированными с результативным признаком и поэтому входят в уравнение регрессии на первых шагах. Поскольку эти главные компоненты поддаются содержательной интерпретации, то появляется возможность интерпретации и уравнения регрессии, причем в терминах главных компонент, а не только исходных признаков. Последние главные компоненты содержат несущественные (второстепенные) сведения. Поэтому когда уравнение строится на наиболее весомые главные компоненты, его устойчивость повышается, а это значит, что изменение состава выборки не оказывает существенного влияния на главные компоненты, и, следовательно, на коэффициенты регрессии.

Отказ от части информации приводит к ухудшению МНК-оценок. Например, может уменьшиться множественный коэффициент детерминации или может возрасти средняя ожидаемая ошибка прогноза. Однако повышение устойчивости уравнения вполне компенсирует названные потери. Этим объясняется расширение использования главных компонент при построении уравнения регрессии в проводимых исследованиях.

Может возникнуть ситуация, когда полученные модели, с точки зрения решаемых задач, представляются исследователю неадекватными. Это происходит, например, из-за наличия аномальных наблюдений или отклонения распределения от многомерного нормального, что, в свою очередь, приводит к необходимости использования нелинейных моделей. [6]

Вообще требование интерпретируемости результатов (сложная и не формализуемая задача) накладывает серьезные ограничения на класс допустимых преобразований и зачастую приводит к значительной корректировке плана исследования по ходу анализа. Поэтому исследователь должен четко представлять возможности, область применения, преимущества и недостатки каждого метода, так как некорректное применение одного из них может дискредитировать сам метод.

Полученные выводы могут быть использованы при выработке рекомендаций для разных групп объектов.

6. Заключение

Итак, исследование начинается с наиболее простых методов, но если их результаты не устраивают пользователя, применяется более тонкая методика. Поскольку в работе используется несколько математико-статистических методов, исследователь должен выстроить (и логически обосновать) последовательность применения этих методов, чтобы они не конкурировали, а дополняли друг друга. Необходимо также обосновать правила сравнения результатов, полученных по различным методам. Особое внимание уделяется возможным разветвлениям математического метода и процесса исследования в целом. При наличии нескольких путей решения какой-то из задач (например, проверки значимости влияния входящей в уравнение переменной) выбирается наиболее эффективный способ, причем выбор этот аргументируется. Если же возникла ситуация, когда полученные результаты совсем не удовлетворяют исследователя, то можно попытаться повысить адекватность модели путем перехода от линейных методов к нелинейным. (Например, метод Бокса-Кокса позволяет с помощью нелинейного преобразования перейти к новому набору переменных так, чтобы при этом максимально возрос множественный коэффициент детерминации).

Из-за специфики реальных задач трудно дать единую схему исследования, пригодную во всех случаях. Как правило, заранее нельзя указать метод, который приведет к наилучшим результатам. Тем не менее, некоторые рекомендации могут быть полезны. Например, при всех недостатках матричного алгоритма регрессионного анализа (множественная регрессия), возможно, имеет смысл начать именно с этой схемы. Не столько для того, чтобы получить само уравнение регрессии, сколько для получения множественного коэффициента детерминации. Он будет максимальным среди возможных. Поэтому его можно использовать как некоторый эталон и оценивать качество всех остальных уравнений еще и по разности коэффициентов множественной детерминации, которая покажет долю утерянной информации.

Пошаговые схемы «включения» и «включения-исключения» целесообразно использовать параллельно. До какого-то шага результаты будут совпадать, а дальше станут различаться из-за того, что некоторые переменные будут выбрасываться из уравнения регрессии. Нарушение устойчивости уравнения показывает сомнительную целесообразность продолжения включения новых признаков в уравнение. Возможно, следует изменить уровень значимости.

При сравнении результатов пошаговых процедур и регрессии на главные компоненты следует учитывать простоту интерпретации пошагового уравнения и более высокую устойчивость уравнения регрессии на главные компоненты, а также различия МНК-оценок – причем не столько абсолютные, сколько относительные. Для

каждого сравниваемого уравнения целесообразно рассчитать таблицу остатков и сравнить эти остатки. При этом также играют роль не только абсолютные значения остатков, но и относительные, что поможет уточнить область применения результатов и выводов.

Составляя план исследования, корректируя его в процессе получения промежуточных результатов и их содержательной интерпретации, комбинировать методы следует так, чтобы они не конкурировали, а дополняли друг друга. Аналогично следует относиться и к получаемым при этом результатам.

Изложенное показывает, что в реальных исследованиях редко встречается ситуация, когда заранее можно указать однозначную «линейную» последовательность применяемых методов. Значительно чаще используемая методика исследования может быть представлена в виде «дерева». И выбор направления продолжения исследования «в узлах» такого «дерева» зависит от промежуточных результатов. Это затрудняет «выдачу рецептов на все случаи жизни». Однако, опираясь на сформулированные принципы корректного применения методов, исследователь может решить поставленную перед ним задачу.

Литература

1. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. – М.: ЮНИТИ, 1998.
2. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. – М.: Финансы и статистика, т.1 – 1986, т.2 – 1987.
3. Дубров А.М. Компонентный анализ и эффективность в экономике. – М.: Финансы и статистика, 2002.
4. Ниворожкина Л.И., Арженовский С.Б. Многомерные статистические методы в экономике. – М.: Дашков и Ко, 2009.
5. Плис А.И., Сливина Н.А. MATHCAD: математический практикум для инженеров и экономистов. – М.: Финансы и статистика, 2003.
6. Симчера В.М. Методы многомерного анализа статистических данных. – М.: Финансы и статистика, 2008.

References

1. Aivazian S.A., Mhitarian V.S. Applied statistics and essentials of econometrics. – M.: UNITY, 1998.
2. Dreyper N., Smith G. Applied regression analysis. – M.: Finance and statistics, t.1 – 1986, t.2 – 1987.
3. Dubrov A.M. Component analysis and efficiency of economics. – M.: Finance and statistics, 2002.
4. Nivorozhkina L.I., Arzhenovsky S.B. Multivariate statistical methods in economics. – M.: Finance and statistics, 2009.
5. Plis A.I., Slivina N.A. MATHCAD: mathematical practical work for engineers and economists. – M.: Finance and statistics, 2003.
6. Simchera V.M. Methods of multivariate analysis of statistical data. – M.: Finance and statistics, 2008.