П.С.Ростовцев, В.С.Костин, А.Л. Олех, А.С. Жданов

Новосибирский государственный университет
Ул. Пирогова, 2 Новосибирск, 630090
Институт Экономики и организации
промышленного производства СО РАН
пр. акад. Лаврентьева, 17, Новосибирск, 630090, Россия

АВТОМАТИЗАЦИЯ АНАЛИЗА СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ДАННЫХ. ДЕТЕРМИНАЦИЯ МОДЕЛЕЙ[1]

Введение

Можно различными путями подходить к анализу данных социально-экономических исследований.

Первый путь - построить теоретическую картину явления и проверить заранее сформулированные модели на эмпирических данных. Но жизнь богаче любой теории, а теория должна иметь свои эмпирические основания.

Второй путь - получить соответствующую теме исследования информацию и "ловить рыбку" из этих данных: изучать разнообразные взаимосвязи, выделять выборки интересующие совокупности объектов, испытывать на них различные модели явлений, глядишь, сами данные наведут на закономерности, которые можно увязать в теорию.

Этот "эмпирический" подход чреват ошибками, так как данные могут иметь случайные отклонения, которые легко принять за закономерности. Руководствуясь статистическими критериями - проверяя гипотезы независимости, гипотезы об отклонении параметров от их ожидаемых значений, число таких ошибок можно существенно сократить. В соответствии с теорией проверки гипотез, гипотезы отвергаются, если получены маловероятные значения критериев. Порог для этой вероятности - "уровень значимости" - назначается заранее. Чаще всего отвергнутая гипотеза означает, что возможна взаимосвязь переменных, наличие регрессионной зависимости и др. Традиционно уровень значимости назначается равным $\alpha=0.05$ или $\alpha=0.01$.

Предположим, наши (анкетные) данные получены случайным образом, так, что все статистики критерия проверяемых нами гипотез независимы - ленивый интервьюер заполнял анкеты сидя под кустиком. Пусть мы проверяем 100 гипотез (к примеру, о независимости переменных). Тогда при уровне значимости $\alpha=0.05$ мы с вероятностью $1-(1-0.05)^{100}=0.994$, а при $\alpha=0.01$ - вероятностью $1-(1-0.01)^{100}=0.634$, мы получим статистики критерия, которые, казалось бы, стоит интерпретировать, несмотря на то, что данные идеально плохи. Это проблема множественных сравнений.

Что же, правы "теоретики"?

Автоматизация анализа данных, которой мы занимались ранее [8], была преимущественно направленным перебором коэффициентов взаимосвязи, поиском связанных переменных. По причине множественных сравнений результат мог получиться сомнительным, хотя, конечно, если существует закономерность, она имеет больший шанс быть обнаруженной, чем ложная закономерность. Такое можно сказать о множестве методов, которые ведут перебор "информативных" признаков. Эта автоматизация вполне соответствует эмпирическому подходу.

В данной работе мы надеемся в определенной степени реабилитировать эмпирический подход, представив процедуру детерминации групп объектов, отличающихся моделями данных. В этой процедуре значимость статистик определяется с учетом множественного сравнения на основе компьютерного эксперимента, имитирующего данные в условиях независимости.

[1] *Исследование поддержано грантом РФФИ 00-06-80221*

А именно, рассматривается одновременно множество пар "группа объектов - модель". Для каждой пары изучается, отличается ли модель данных внутри группы объектов от модели для совокупности объектов вне этой группы. Нулевая гипотеза - это гипотеза неразличимости моделей в группах и их дополнениях. Критерий - вероятность получить случайно в условиях гипотезы хотя бы одно сочетание группы объектов и модели, лучше подчеркивающее индивидуальность группы. Слова "хотя бы одно сочетание" имеет в методе ключевое значение: обычно, за исключением специальных случаев статистического анализа, чаще дисперсионного анализа [2, 3, 12, 18], модели связи переменных рассматриваются без их сопоставления между собой.

Метод является развитием работ [9, 10], в которых ранее удалось реализовать метод множественных сравнений в анализе таблиц для неальтернативных вопросов, где были представлены основные идеи множественных сравнений для простых статистик в детерминационном анализе и в типологическом группировании, а также работы [7], в которой впервые было реализовано множественное сравнение при анализе связи непересекающихся групп объектов с проблемной группой.

Метод реализован в виде компьютерной программы, которая совместима по данным со статистическим пакетом SPSS [17] и может быть подключена к нему пунктом меню. Работа ориентирована на приложения в социологических исследованиях, но статистические методы универсальны, и она может пригодиться в различных областях.

Вклад соавторов. Метод в основном был разработан П.С Ростовцевым, им же был подготовлен текст работы. Программная реализация метода сделана в основном В.С.Костиным, глубокое понимание сути проблемы и статистических вопросов реализации метода, позволили уточнить ряд его важных деталей. В частности, идея использования аппроксимации множественной значимости с помощью бета распределения ***B(1,b,x)*** является результатом совместного труда Костина и Ростовцева. Возможность использования неальтернативных признаков для задания групп объектов обеспечил А.С.Жданов, управление форматом выдачи - А.Л.Олех.

Описание групп объектов

В отличие от обычных методов анализа данных, группы объектов, которые мы хотим рассматривать в данной работе не обязательно должны разбивать совокупность объектов на непересекающиеся подмножества. Для того чтобы не утомлять читателя формальными способами описания групп, мы рассмотрим их на примерах.

Естественный путь описания группы объектов - задание ее в виде комбинаций значений. В простейшем случае можно просто указать значения неколичественной переменной, например, таким образом можно отметить группы женатых/замужних, не состоявших в браке и разведенных (схема 1).

Схема 1. Группы по категориальной переменной

Семейное положение	√	Женат
	√	Холост
		Вдов
	√	Разведен

Несколько сложнее задание групп в виде комбинации значений переменных (схема 2).

Схема 2. Группы по сочетанию категориальных переменных

Семейное положение	√	Женат	Пол	√	Мужчины
				√	Женщины
	√	Холост			
		Вдов			
		Разведен	Возраст	√	Молодые
				√	Старые

В последнем примере рассматриваются группы семейных, женатых мужчин, замужних женщин, не состоявших в браке, молодых и старых разведенных. Заметим, что

рассматриваются одновременно группы семейных, женатых мужчин и замужних женщин, хотя две последних группы являются частями группы семейных.

В социологии, в прикладных анкетных исследованиях очень часто используются неальтернативные[2] вопросы (вопросы, позволяющие дать одновременно несколько ответов). Они тоже могут быть использованы для формирования групп в такой же иерархической структуре. Например, группы семей по ответам на вопрос "Имеете ли Вы следующие дорогостоящие предметы недвижимость?" и по виду жилья (схема 3.).

Группы могут быть, также заданы в виде объединения нескольких подобных схем.

Схема 3. Группы по сочетанию ответов на неальтернативный и альтернативный вопрос

Собственность		Автомобиль	Жилье	√	Квартира
				√	Собственный дом
		Дача	Жилье	√	Квартира
				√	Собственный дом
		Компьютер	Жилье	√	Квартира
				√	Собственный дом

Как мы видим из представленных примеров, могут быть заданы как пересекающиеся, так непересекающимися группы, они могут быть даже включены в друг друга.

Современное математическое обеспечение статистического анализа позволяет предварительную подготовку данных, поэтому группы объектов могут быть получены произвольными логическими конструкциями из количественных и качественных переменных.

Группы и дихотомические переменные

Характеристики группы объектов нет смысла рассматривать изолированно, всегда необходимо их сравнивать с характеристиками всей совокупности объектов, с собственными представлениями об этих характеристиках, с дополнением группы до всей совокупности объектов. Например, является ли средний возраст 33 года работников некоторой сфере занятости особенной характеристикой или нет, можно понять, только сопоставив с данными в других сферах занятости. Для автоматизации анализа данных целесообразно такое сравнение проводить с дополнением группы, т.е. статистические характеристики рассматривать для градаций соответствующей дихотомической переменной.

Задавая группы, мы одновременно задаем множество дихотомических переменных, которые используются для оценки особенностей моделей на этих группах.

Модели сравнения групп.

Заданные группы могут быть интересны с самых различных точек зрения.

Например, можно рассмотреть задачи, существенно ли чаще/реже и в большем или меньшем объеме, чем остальные семьи, потребляют спиртное в группах семей, описываемых различными характеристиками (наличием собственности, жилищными условиями и др.). Можно оценить смещения этих групп по составу семей; выяснить насколько они отличаются по распределению времени досуга, а возможно и особенностями взаимосвязи досуга с уровнем заработков.

Для решения этих задач мы рассматриваем группы объектов и их дополнения, исследуя значимость различия моделей в этих совокупностях. Вначале представим модели сравнения групп объектов, отвлекаясь от проблемы множественных сравнений. Поэтому, будем пока

[2] Каждый неальтернативный (многозначный) вопрос в данных хранится обычно в нескольких переменных - либо дихотомических, соответствующих ответам, либо в переменных, содержащих номера возможных ответов. Группы переменных, соответствующие неальтернативным вопросам, мы будем называть групповыми переменными. В группу можно объединить также переменные, значения которых можно трактовать как ответы на многозначный вопрос.

считать, что имеется только одна тестируемая группа наблюдений ***A***, и нужно изучить, чем она отличается от ее дополнения $\overline{A}$. В методах сравнения двух групп объектов, рассмотренных ниже нет ничего нового, но статистики, возникающие в известных методах, составляют базу для решения проблемы выбора значимых пар "модель-группа" в условиях множественных сравнений.

Модель проблемной группы

Пусть имеется группа наблюдений ***B***, которая с какой-то точки зрения может быть проблемной. Например, изучается группа семей, отличающихся бедностью, или школьников, поражающих своим талантом, или покупателей определенного товара в маркетинговых исследованиях. По отношению к этой группе обычно возникают вопросы, в чем причина проблемной характеристики этой группы, какие характеристики объектов сопутствует этой группе, какие нет. Первый шаг необходимого анализа - установление связи тестируемой группы объектов ***A*** с проблемной ***B***. Это достигается исследованием связи двух дихотомических переменных, соответствующих разбиениям совокупности объектов $G=\{A, \overline{A}\}$ и $\{B, \overline{B}\}$.

Существует множество коэффициентов связи между такими переменными, из которых для анализа значимости мы используем статистику $Z=(N_{AB}-E_{AB})/\sigma$ - смещения наблюдаемого числа объектов N_{AB} в пересечении групп ***A*** и ***B*** от ожидаемого E_{AB}, где дисперсия σ^2 вычислена исходя из гипергеометрического теоретического распределения N_{AB} [14, 15]**.** Для больших выборок ***Z*** имеет приближенно стандартное нормальное распределение.

Положительные значения ***Z*** свидетельствуют о положительной связи - $N_{AB}/N_{A.}>N_{.B}/N$ - если объект характеризуется свойством ***A***, то он, вероятно, имеет большую тенденцию характеризоваться свойством ***B***, и наоборот. Отрицательные смещения говорят об отрицательной связи. Наблюдаемый уровень значимости, вероятность $\alpha=P\{|Z|>|Z_{выборочное}|\}$ случайно на независимых данных получить большее по абсолютной величине значение этой статистики, чем выборочное, позволяет понять, не является ли полученное отклонение игрой случая.

Малая вероятность α говорит о том, что по отношению к проблемной группе ***B*** тестируемая группа является "особенной", и ее, возможно, стоит внимательно изучить.

Модель особенного распределения номинальной переменной.

Пусть имеется номинальная переменная ***X***, имеющая ***m*** значений. Например, при исследовании электорального поведения тестируются различные группы по переменной "приверженность различным политическим движениям". Модель проблемной группы - частный случай модели номинальной переменной.

Для выяснения, существенно ли отличается распределение ***X*** в группе ***A***, рассматривается таблица $||N_{ij}||$ сопряженности ***X*** и дихотомической переменной $G=\{A, \overline{A}\}$, где индекс ***i*** соответствует значениям ***G*** (***A*** и $\overline{A}$), а ***j*** - значениям ***X***.

Для оценки значимости различий вычисляется статистика

$L^2 = 2\sum_{i,j} N_{ij} \ln(\frac{N_{ij}}{E_{ij}})$, где E_{ij} - частоты, ожидаемые в условиях независимости ***G*** и ***X***.

Статистика L^2 в условиях гипотезы независимости имеет распределение, близкое к распределению хи-квадрат. Эта статистика предпочтительнее стандартного критерия хи-квадрат, поскольку имеет более точную асимптотику [1]. Также как для статистики **Z**, для выяснения существенности отличия распределения ***X*** в группе ***A*** используется наблюдаемый уровень значимости $\alpha=P\{L^2>L^2_{выборочное}\}$.

В этой модели можно использовать и ранговые, и сгруппированные количественные переменные.

Модель выдающегося среднего

Отличается ли в группе ***A*** средний возраст, результаты психологического тестирования по одному тесту и другие количественные характеристики, имеющие не

слишком скошенное распределение? Для ответа на подобные вопросы мы используем стандартный параметрический подход - критерий Стьюдента [4].

Для проверки значимости различия средних в группах A и $\overline{A}$ в предположении теоретического нормального распределения, используется статистика $t = \frac{\left(\overline{X}_A - \overline{X}_{\overline{A}}\right)}{\sqrt{S_A^2/N_A + S_{\overline{A}}^2/N_{\overline{A}}}}$, имеющая распределение Стьюдента с числом степеней свободы, зависящем от оценок дисперсии S_A^2, $S_{\overline{A}}^2$ и от объемов групп. Измерителем значимости "выдающегося среднего" здесь также является наблюдаемая значимость $\alpha=P\{|t|>|t_{выборочное}|\}$..

Модель выдающегося среднего ранга

Если нас интересует различие в иерархии объектов (иерархия в доходах, в имущественном положении, в результатах тестирования интеллекта и т.п.), то следует воспользоваться статистикой Манна-Уитни [15]. Пусть r_i - ранг i -того объекта по переменной X, причем для групп объектов с одинаковыми значениями r_i ранг определяется как средний ранг этой группы. Тогда статистика Манна-Уитни имеет вид $M\_W=\sum_{i\in A} r_i$. Математическое ожидание ее в условиях одинакового распределения X в группе и дополнении равно $M(M\_W)=1/2*N_A*(N+1)$. Дисперсия в этих условиях имеет вид $D(M\_W)=$ $\frac{N_A * N_{\overline{A}}}{12}\left[N+1-\frac{\sum_l^m N_l(N_l^2-1)}{N(N-1)}\right]$, где m - число значений тестируемой переменной, N_l - частота l-го значения. Статистика M_W - является суммой рангов, однако, если ее поделить на число объектов в группе A (по существу - на константу) она станет средним рангом. Поэтому M_W - характеристика отклонения среднего ранга от его ожидаемого в условиях независимости значения. Асимптотически нормальная в условиях гипотезы совпадения распределений X в A и $\overline{A}$ статистика - $Z=(M\_W-M(M\_W))/(D(M\_W))^{1/2}$ позволяет оценить значимость $\alpha=P\{|Z|>|Z_{выборочное}|\}$ выдающегося среднего ранга.

Модель особенной формы распределения

Модели "выдающегося" среднего и среднего ранга не улавливают изменения формы распределения: t-тест рассчитан на сравнение средних нормальных распределений, тест Манна-Уитни - лучше всего ведет себя при альтернативной гипотезе сдвига распределений. В то же время нередко существенным является именно исследование различий в формах распределения. Например, более растянутая форма распределения по доходам свидетельствует о высокой степени неравенства распределения доходов в обществе. Особенности формы распределения уловит тест Колмогорова-Смирнова [4]:

$K\_S = \sqrt{\frac{N_A N_{\overline{A}}}{N_A + N_A}} \max_x \left|F_A^*(x) - F_{\overline{A}}^*(x)\right|$, где $F_A^*(x)$ и $F_{\overline{A}}^*(x)$ - эмпирические функции распределения в группе объектов A и $\overline{A}$. Группа A будет иметь "особенное" распределение по X, если наблюдаемый уровень значимости $\alpha=P\{K\_S>K\_S_{выборочное}\}$ мал.

Выдающийся вектор средних

Зачастую социологи, психологи, специалисты по маркетингу выбирают для своего исследования достаточно обширный список показателей. Например, специалист, занимающийся продвижением на рынок молочных продуктов, выясняет объемы потребления молока, кисломолочных продуктов, сметаны, сыра и т.п. Желательно иметь картину потребления продуктов в целом и тестировать группы населения, по потреблению всего списка товаров.

Пусть имеется вектор $X=(X_1,\ldots,X_m)^T$.

$$S = \frac{1}{N_A + N_{\bar{A}} - 2}((N_A - 1)S_A + (N_{\bar{A}} - 1)S_{\bar{A}})$$

где S_A и $S_{\bar{A}}$ - выборочные ковариационные матрицы X в A и $\bar{A}$. Для обнаружения существенности отклонения вектора средних в группе А мы используем статистку Хотеллинга [4], которая имеет вид

$$T^2 = \frac{N_A N_{\bar{A}}}{N_A + N_{\bar{A}}}(\bar{X}_A - \bar{X}_{\bar{A}})^T S^{-1}(\bar{X}_A - \bar{X}_{\bar{A}})$$

В условиях теоретической многомерной нормальности, равенства матожиданий и ковариационных матриц статистика

$$F = \frac{N_A + N_{\bar{A}} - m - 1}{(N_A + N_{\bar{A}} - 2)m}T^2$$

имеет распределение Фишера $F(m, N_A + N_{\bar{A}} - m - 1)$. Это позволяет вычислить искомую значимость $\alpha = P\{|T^2| > |T^2_{выборочное}|\}$.

Модель различия регрессионных зависимостей

Предположим, мы строим регрессионную зависимость вектора переменных $X=(X_1,\ldots,X_m)^T$ на вектор Y. Очень может быть, что на совокупности А будут получены одни коэффициенты, а на совокупности $\bar{A}$ - другие. Например, затраты рабочего времени в городе иначе связаны с доходами, чем в деревне.

Для обнаружения этого мы пользуемся подходом Чоу [13], который заключается в следующем.

Вместо исходных m рассматривается $2m$ независимых переменных, включающих множество переменных X^A, совпадающих с X на группе объектов и равных нулю на ее дополнении, и множество переменных $X^{\bar{A}}$, равных нулю на объектах группы и совпадающих с X на дополнении. Т.е. решается уравнение

$$\begin{pmatrix} X^A & 0 \\ 0 & X^{\bar{A}} \end{pmatrix}\begin{pmatrix} B^A \\ B^{\bar{A}} \end{pmatrix} = \begin{pmatrix} Y^A \\ Y^{\bar{A}} \end{pmatrix},$$

где Y^A и $Y^{\bar{A}}$ - значения Y для группы и ее дополнения соответственно. Для выявления значимости отличия регрессии нужно проверить гипотезу о равенстве векторов коэффициентов этой объединенной функции регрессии B^A и $B^{\bar{A}}$. Коэффициенты находятся обычным методом наименьших квадратов в предположении, что соблюдаются все условия классической линейной модели регрессии.

В соответствии с методом проверки гипотез о линейных комбинациях коэффициентов регрессии, описанным в, статистика для оценки различия моделей имеет вид

$$F = \frac{1}{ms^2}(B^A - B^{\bar{A}})^T\left((X^A X^{AT})^{-1} + (X^{\bar{A}} X^{\bar{A}T})^{-1}\right)^{-1}(B^A - B^{\bar{A}}),$$

где s - стандартная ошибка регрессии. Если верна гипотеза о совпадении коэффициентов, эта статистика имеет распределение Фишера $F(m,N-2m)$. Таким образом, наблюдаемая значимость группы A оценивается величиной $\alpha = P\{F > F_{выборочное}\}$.

Значимость как единый измеритель сравнения моделей

Указанный список можно продолжать бесконечно, необходимо лишь для каждой новой модели предложить статистику сравнения моделей в группе и дополнении, а также способ вычисления значимости. Разнообразные схемы оценки качества моделей сводятся к одной шкале - шкале вероятности случайно получить лучшие модели, в условиях однородности данных. Заметим, что обычно при исследовании взаимосвязей оказывается, что чем больше значение статистики, тем надежнее выявлена связь. Для шкалы наблюдаемого уровня значимости - наоборот, чем меньше уровень значимости, тем это лучше для выявления закономерности. Поэтому мы говорим о большей значимости, если оценки величина наблюдаемой значимости меньше.

Есть одно неудобство в использовании наблюдаемой значимости - это то, что, при достаточно большой связи переменных, она близка к нулю. Например, в модели проблемной группы наблюдаемые значимости $Z=5$ и $Z=6$ равны $5.7*10^{-07}$ и $2*10^{-09}$. Использование логарифма значимости или записи в виде мантиссы и степени не наглядно. В работе [6] предложено в качестве универсальной шкалы использовать Z-статистики, вычисляемые как квантиль порядка $p=1-\alpha/2$ нормального распределения ($Z=\Phi^{-1}(1-\alpha/2)$), где α - наблюдаемая значимость статистики модели. В ряде моделей (в нашем случае - в моделях проблемной группы, среднего ранга и сравнения средних) Z-статистика трактуется как число стандартных отклонений, на которое отклонилось значение статистики от ожидаемого значения. В других моделях значение Z - сопоставимое обобщение этой шкалы.

Наблюдаемая значимость как случайная величина

Пусть $\alpha(G,M)$ - наблюдаемый уровень значимости сходства моделей на группах, определяемых дихотомией G . Это означает, что вероятность получить некоторую статистику для этого сочетания, большую чем ее выборочное значение равна $\alpha(G,M)$, (к примеру, $\alpha=\alpha(G,M)=P\{|Z|>|Z_{выборочное}|\}$ для модели проблемной группы). Если проводить эксперименты на некоторых случайных данных (например, сгенерированных на компьютере), величина $\alpha(G,M)$ будет случайной величиной. Несложно показать, что в условиях гипотезы однородности групп относительно модели она будет иметь равномерное на отрезке $(0,1)$ распределение. Точнее, распределение α аппроксимируется равномерным распределением в моделях, использующих стандартную аппроксимацию (нормальную, хи-квадрат, Z-Колмогорова-Смирнова) и действительно равномерное в практически недостижимых условиях нормальности распределения данных или регрессионных остатков.

Множественная значимость групп и моделей

Как было сказано выше, пользуясь критерием значимости, на множестве пар (модель - группа объектов), мы можем с высокой вероятностью получить хотя бы одну "значимую" статистику, даже если группы объектов и их дополнения однородны по отношению к этим моделям. Это не должно удовлетворять добросовестного исследователя, заботящегося о достоверности выводов. Поэтому разумным является критерий, который в условиях однородности практически не предлагает случайно для анализа ни одной ошибочной закономерности.

Пусть тестируется k сочетаний $(G_1,M_1),(G_2,M_2),\ldots,(G_k,M_k)$ групп и моделей. Пусть задана α_k - вероятность ошибочно признать различие модели на группе и на ее дополнении *хотя бы для одного сочетания* модели и группы. Величину α_k будем называть уровнем множественной значимости.

Если бы все модели на группах и их дополнениях были независимы, то для достижения этой вероятности достаточно для каждого сочетания (G,M) проверять, будет ли наблюдаемая значимость $\alpha(G,M)<1-(1-\alpha_k)^{1/k}$. Однако эта оценка будет неверна, если существует взаимосвязь моделей. В частности, если сочетания моделей и групп дублируют друг друга, в качестве критического уровня значимости вместо $1-(1-\alpha_k)^{1/k}$ следует взять непосредственно α_k. Универсальный метод Бонферрони [3] использует в качестве критического значения уровня значимости α_k/k, однако он использует верхнюю оценку вероятности ошибки и поэтому груб.

Гипотеза об отсутствии связи моделей и групп.

Для определения значимости сочетаний групп и моделей с точки зрения множественных сравнений мы воспользуемся гипотезой об отсутствии связи дихотомических переменных $G_1,\ldots,G_k$, соответствующих группам, с переменными, участвующими в построении моделей $M_1,\ldots,M_k$ (а значит и однородности групп по отношению к моделям). Для простоты будем называть эту гипотезу гипотезой независимости групп и моделей.

Как было отвечено выше, наблюдаемая значимость является универсальным показателем неоднородности групп для всех рассматриваемых нами моделей. Этим показателем мы

будем пользоваться для определения множественной значимости сочетаний ***(G,M)***. Для этого мы определим распределение величины $\alpha_{min}=\min_i \alpha(G_i, M_i)$ в условиях гипотезы независимости групп и моделей. Квантиль β порядка α_k для α_{min} мы будем считать критическим значением для статистик $\alpha(G_i,M_i)$, $i=1,\ldots,k$. Это означает что, если $\alpha(G_i,M_i)<\beta$, то модель M_i на группе G_i существенно отличается от модели M_i на дополнении этой группы. Вероятность случайно получить значимую статистику для хотя бы одного сочетания группы и модели будет равна нашему заранее выбранному уровню значимости α_k, т.е. $P(\bigcap_{i=1}^{k}\{\alpha(G_i, M_i) < \beta\}) = \alpha_k$. Это значит, что в зависимости от выбранной модели сравнения, группа отличается средним, формой распределения некоторой переменной, коэффициентами регрессионного уравнения и т.д., и такое отличие можно получить, перебрав все заранее отобранные сочетания моделей и групп, лишь с вероятностью α_k.

Наблюдаемая множественная значимость

Для пары группа-модель ***(G,M)*** наблюдаемая множественная значимость определяется как вероятность случайно, в условиях гипотезы независимости групп и моделей, получить хотя бы одну более значимую пару ***(G,M)***. Иными словами величина $\beta_{наблюдаемае}$ определяется по формуле $\beta_{наблюдаемае}=P\{\alpha_{min}<\alpha(G,M)\}$.

Оценивание множественной значимости. Статистический эксперимент

Таким образом, суть проблемы состоит в определении распределения α_{min} - минимального в условиях гипотезы независимости групп и моделей уровня значимости.

Поскольку $\alpha(G,M)$ для любой модели имеет равномерное распределение, в условиях независимости всех k пар (G_i,M_i) функция распределения α_{min} равна $F_{\alpha_{\min}}(x) = 1-(1-x)^k$. В условиях их полной взаимосвязи (дублирования) - $F_{\alpha_{\min}}(x) = 1-(1-x) = x$. В общем случае таких простых формул у нас нет, и мы вынуждены прибегнуть к статистическому эксперименту, в котором имитируется независимость моделей и групп.

Напомним, что наши данные состоят из дихотомических переменных - индикаторов групп объектов и переменных, предназначенных для построения моделей. Смысл эксперимента состоит в перемешивании данных по переменным - индикаторам групп. За счет перестановки данных сведения о группах могут быть приписаны произвольному объекту.

Производя $N_э$ экспериментов, мы в q-том эксперименте вычисляем значения $\alpha_{q,i}=\alpha(G_i,M_i)$, $q=1,\ldots,N_э$, , а также минимальное значение этой величины $\alpha_{\min}^{q}$.(см. схему 4).

Схема 4. Схема результатов экспериментов

Номер эксперимента	$\alpha(G_1,M_1)$	…	$\alpha(G_k,M_k)$	α_{min}
1	$\alpha_{1,1}$	…	$\alpha_{1,k}$	$\alpha_{\min}^{1}$
…	…	…	…	…
$N_э$	$\alpha_{Nэ,1}$	…	$\alpha_{Nэ,k}$	$\alpha_{\min}^{Nэ}$

Таким образом получается выборочное распределение α_{min}.

Пусть $\alpha_{\min}^{(1)},\ldots,\alpha_{\min}^{(q)},\ldots,\alpha_{\min}^{(N_э)}$ - упорядоченный (статистический) ряд наблюдений α_{min}, полученных в экспериментах, тогда величина $\alpha_{\min}^{(q)}$ может служить оценкой квантили порядка $q/N_э$. Оценка квантили порядка β этого распределения дает критическое значение для определения значимых сочетаний типа "группа-модель" (G_i,M_i).

Оценка наблюдаемой множественной значимости для $\alpha=\alpha(G,M)$ состоит в поиске значений $\alpha_{\min}^{(q)}$, близких по величине к α. Этой оценкой будет величина $\beta_{наблюдаемое}=q/N_э$.

Аппроксимация распределения α_{min}

Часто значение α, полученное на эмпирических данных, выходит далеко за границы разброса статистического ряда $\alpha_{\min}^{(q)}$. Тогда непосредственная оценка наблюдаемой значимости на основе этого ряда невозможна. Не может удовлетворить, также, ее дискретность, также как дискретность значений, выбираемых в качестве критических значений $\alpha=\alpha(G,M)$. Поэтому необходима аппроксимация распределения α_{min}.

Для построения такой аппроксимации вспомним, что в условиях независимости функция распределения α_{min} имеет вид $F_{\alpha_{\min}}(x) = 1-(1-x)^k$. Обобщая эту формулу, мы предполагаем, что эта функция имеет вид $F_{\alpha_{\min}}(x) = 1-(1-x)^b$, где коэффициент b интерпретируется как некоторое, скрытое в наших данных, число степеней свободы. Это ничто иное, как функция бета распределения $B(1,b,x)$ [11][3].

Для оценки параметра b достаточно использовать среднее имеющихся выборочных значений: $\hat{b} = \frac{1}{\alpha_{\min}} - 1$. Оценка имеет смещение порядка **1/Nэ**, которое может быть скорректировано до порядка $1/N_э^2$.

С использованием данной аппроксимации критическое значение для $\alpha=\alpha(G,M)$ оценивается как $1-(1-\alpha_k)^{1/b}$, а наблюдаемая множественная значимость α - как $\beta_{наблюдаемое} = B(1,b,x)=1-(1-\alpha)^b$.

Неприятности, их преодоление

Теоретически все выглядит прекрасно, но на практике следует учесть ряд особенностей данных и используемых показателей.

Прежде всего, это то, что для оценки значимости моделей проблемных групп, номинального распределения, ранговых критериев Вилкоксона и Колмогорова-Смирнова привлекаются непрерывные аппроксимации функций распределения соответствующих статистик. Фактически же, на малых группах эти статистики дискретны.

Нулевая гипотеза для проверки равенства матожиданий (модели выдающегося среднего и вектора средних) предполагает нормальность распределения. При проведении массовых расчетов мы не имеем возможности проверки нормальности. В равной степени это касается и распределения регрессионных остатков.

Так что в реальности предполагаемые законы распределения могут нарушаться, причем распределение одного и того же критерия может быть различным для различных групп объектов. Поэтому в реализации метода предусмотрена корректировка оценок наблюдаемых значимостей. Это возможно благодаря статистическим экспериментам, которые дают эмпирические распределения оцененных значимостей для каждого сочетания "группа-модель". Теоретически это распределение должно быть равномерным, но неточность аппроксимации вносит искажения. По сути, здесь происходит "исправление" искаженного равномерного распределения. Для этого мы применяем бета распределение.

Пример использования

Использование метода и программы мы проиллюстрируем на данных 1998 г. Российского мониторинга экономического положения и здоровья населения (RLMS, [5], используется только городское население). В качестве примера изучим, как связано с

[3] Ранее, в работах [9, 10], мы делали аналогичную аппроксимацию с помощью функции бета распределения $B(a,b,x)$, в которой оценивались оба параметра, но опыт нас убедил в том, что аппроксимация вида $B(a,b,x)$ не хуже, а с точки зрения устойчивости даже значительно лучше. Проверка соответствия распределения с помощью критерия Колмогорова-Смирнова на реальных данных показала, пригодность данной аппроксимации. Кроме того были проведены вычислительные эксперименты: генерировались многомерные нормальные случай векторы с нетривиальной ковариационной матрицей, компоненты которых преобразовывались в шкалы значимости. Оказалось, что в этих экспериментах даже на 10000 наблюдений гипотеза о бета $B(1,b,x)$ распределении не отвергается. Это позволило нам принять решении об использовании данной аппроксимации.

материальным состоянием семей потребление спиртных напитков и табачных изделий. Цель примера - иллюстрация метода, а не проведение глубокого содержательного исследования.

Может быть, спиртные напитки и табак употребляют, поскольку материальное положение позволяет? Может быть, их употребление вызывает в целом падение уровня достатка? Возможно, среди употребляющих спиртные напитки иные связи между материальным благосостоянием и доходами? А может быть, здесь имеется связь с жилищными условиями?

В используемых данных по городским домашним хозяйствам RLMS имеется информация о покупках 2600 семей, сделанных в течение одной недели (молочных продуктов, спиртного и табака, сладостей и другого), а также жилая площадь, денежные доходы населения, сведения об имуществе (наличие автомобиля, дачи, компьютера и т.п.) полезные для поставленной задачи.

В качестве тестируемых групп нами взяты ответы о покупке в последнюю неделю спиртного и табака: группы потребителей водки, вина, пива и табака, а также группа семей, не покупавшая эти товары.

Тестирование будем проводить по моделям

- Модель проблемной группы: изучим, кто не смог ответить по поводу типа жилья, где проживает семья (ответившие "не знаю" (возможно - нет жилья), не потребители ли водки?).
- Модель особенного номинального распределения: проверим, различаются ли распределения по типу жилья в указанных группах (отдельная квартира, часть квартиры, отдельный дом, часть дома, часть дома) в указанных группах.
- Модель выдающегося среднего ранга: отличаются ли группы по иерархии по площади жилья на одного человека.
- Модель выдающегося среднего: проверим различие групп по среднему логарифму жилплощади.
- Модель особенного вектора средних: одновременно проверим отличие средних логарифмов жилплощади и денежных доходов.
- Модель регрессии: отличаются ли группы различной зависимостью жилплощади от уровня доходов.

Использование показателей доходов и жилой площади, измеренных в логарифмической шкале у нас, как, впрочем, и во многих других исследованиях, объясняется необходимостью получения менее скошенного распределения этих переменных.

Анализ простых наблюдаемых значимостей показал, что при обычном отборе по 5% уровню следует обратить внимание на 23 сочетания моделей и групп (см. таблицу 1).

Оценка множественных значимостей моделей на основе 1000-кратного перемешивании данных, выявила лишь 15 значимых на уровне 5% сочетаний групп и моделей (см. таблицу 2). Восемь, по обычным понятиям, значимых статистик при множественных сравнениях не следует считать надежным свидетельством связи в данных. Например, проблемная группа (семьи, в которых не смогли определить тип жилья) при обычном подходе была бы без сомнения значимой на уровне 0.5% для семей, не покупавших спиртного и табака, однако множественная значимость ее здесь - 14.6%, поэтому такой вывод можно поставить под сомнение.

Таблица 1. Наблюдаемые значимости отличия моделей в группах семей покупателей спиртного и табачных изделий

Модели	Группы				
	Не покупали	Водка	Вино	Пиво	Табак
Проблемная группа - возможно нет жилья	***0.0049***	0.1360	0.0195	***0.0493***	***0.0143***
Номинальное распределение (по типу жилья)	***0.0080***	0.0522	***0.0330***	***0.0001***	0.0973
Средний ранг жилплощади	***0.0006***	0.8000	0.3950	***0.0096***	***0.0000***
Форма распределения (жилплощадь)	***0.0000***	0.5280	0.1970	***0.0078***	***0.0000***
Средний Ln жилплощади	***0.0003***	0.8050	0.5290	***0.0040***	***0.0000***
Средние Ln жилплощади и дохода	***0.0000***	***0.0000***	***0.0000***	***0.0000***	***0.0000***
Регрессия: жилплощадь от дохода -	***0.0000***	0.1760	0.5730	***0.0000***	***0.0000***

В результате экспериментов получен коэффициент $b=32.9$ для распределения $B(1,b,x)$, аппроксимирующего распределение α_{min}. Таким образом, для данной таблицы моделей и групп для вычисления множественной значимости по обычной α можно воспользоваться формулой $\alpha_k=1-(1-\alpha)^{32.9}$.

Связь этой проблемной группы с покупками водки незначима как с точки зрения обычного анализа (наблюдаемый уровень значимости 0.136), так и с точки зрения множественной значимости (наблюдаемый уровень значимости 0.992). В последнем случае это выглядит очевиднее, так как в случае независимости моделей и групп не хуже результаты можно получить с вероятностью 0.992.

Таблица 2. Оценки наблюдаемых множественных значимостей отличия моделей в группах семей покупателей спиртного и табачных изделий

Модели	Группы				
	Не покупали	Водка	Вино	Пиво	Табак
Проблемная группа - возможно нет жилья	0.1460	0.9920	0.4660	0.8170	0.3900
Номинальное распределение (по типу жилья)	0.2320	0.8330	0.6740	***0.0032***	0.9690
Средний ранг жилплощади	***0.0190***	1.0000	1.0000	0.2780	***0.0000***
Форма распределения (жилплощадь)	***0.0000***	1.0000	0.9980	0.2070	***0.0000***
Средний Ln жилплощади	***0.0100***	1.0000	1.0000	0.1290	***0.0001***
Средние Ln жилплощади и дохода	***0.0000***	***0.0000***	***0.0000***	***0.0000***	***0.0000***
Регрессия: жилплощадь от дохода -	***0.0000***	0.9990	1.0000	***0.0000***	***0.0000***

У нас нет возможности изучать все значимые сочетания, обсудим качестве примера наиболее интересные с нашей точки зрения модели. А именно, рассмотрим модели формы распределения, вектора средних и регрессии для группы покупателей табачных изделий. Условно будем называть эту группу "курильщиками" Группа объединяет 1076 семей, дополнение - 1528 семей, однако при получении некоторых моделей часть объектов отбрасывалось из-за наличия в них неопределенных значений.

Модель различия формы распределения

Описательные характеристики для модели Колмогорова-Смирнова находятся в таблице 3. Максимальное различие функций распределения в группе и ее дополнении равна 0.112, причем это - положительная разность; отрицательная разность - 0.018 - незначительна.

Поэтому можно утверждать, что в целом распределение по жилплощади семей курильщиков существенно сдвинуто влево.

Нами получена, также значимость в единицах Z-шкалы нормального распределения Z=6.28. С позиций множественных сравнений обычная наблюдаемая значимость и Z-шкалы представляют собой описательные статистики.

Таблица 3. Статистики теста Колмогорова-Смирнова

Наибольшие разности функций распределения	Абсолютная	0.112
	Положительная	0.112
	Отрицательная	-0.018
Z Колмогорова-Смирнова		2.7
Асимптотическая двусторонняя значимость		0.000

Полезно взглянуть, также, на гистограммы распределения функций переменной в тестируемой группе и в ее дополнении (см. рисунок 1). Кроме некоторого смещения влево гистограммы для курильщиков, заметна существенно большая пикообразность распределения по жилплощади у тех, кто их табак не покупает.

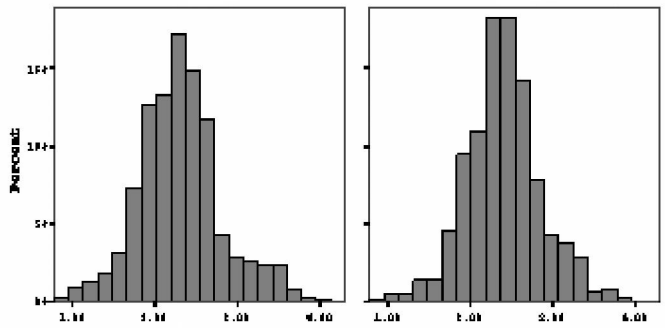

Рисунок 1. Гистограмма распределения в тестируемой группе (слева) и ее дополнении (справа).

Вероятно, жилищные условия семей курильщиков больше варьируют, чем семей, где нет курильщиков.

Модель различия векторов средних

Нашим методом (и критерием Хотеллинга) также обнаружено, что средние векторы средних логарифмов доходов и жилплощади курильщиков также существенно отличаются. Действительно (таблица 4), средний логарифм жилплощади у курильщиков несколько меньше, чем в противоположной группе (2.32 против 2.42), а для логарифма душевого дохода наблюдается обратное соотношение (5.71 против 5.53).

Z-статистики сравнения этих средних равны -4.73 и 4.79 для логарифмов жилплощади и душевых доходов. "Курильщики" несколько богаче некурильщиков, но жилплощадь их несколько меньше.

В целом наблюдаемая значимость различия векторов равна нулю, соответствующее значение Z=6.5.

Таблица 4. Сравнение векторов средних

Покупали табачные изделия?		Логарифм жилплощади	Логарифм душевого дохода	Количество наблюдений
1 Да	Среднее	2.32	5.71	922
	Стд.откл.	0.50	0.90	
2 Нет	Среднее	2.42	5.53	1287
	Стд.откл.	0.45	0.85	
Всего	Среднее	2.38	5.60	2209
	Стд.откл.	0.47	0.88	

Сравнение регрессионных уравнений

Полученная нулевая значимость различия уравнений регрессии (Z=6.37) свидетельствуют, что связь доходов с размерами жилой площади у "курильщиков" и "некурящих" различна.

Таблица 5. Сравнение регрессионных коэффициентов

Группы	Переменные	B	Std.Dev.B	T	Sig
Тестируемая группа	Const	1.321	0.095	13.86	0.0000
-	LN доходов	0.175	0.016	10.60	0.0000
Дополнение	Const	1.700	0.083	20.43	0.0000
-	LN доходов	0.130	0.015	8.72	0.0000
Сравнение	Const	-0.379	0.127	-3.00	0.0030
(разность коэфф.)	LN доходов	0.045	0.022	2.03	0.0428

Таблица 5 показывает детали такого различия. Она содержит описание уравнений регрессии для "курильщиков", для "некурящих", а также статистики сравнения коэффициентов регрессии. С точки зрения множественных сравнений статистики значимости для коэффициентов здесь имеют, конечно, описательный характер, но, по-видимому, без них не обойтись.

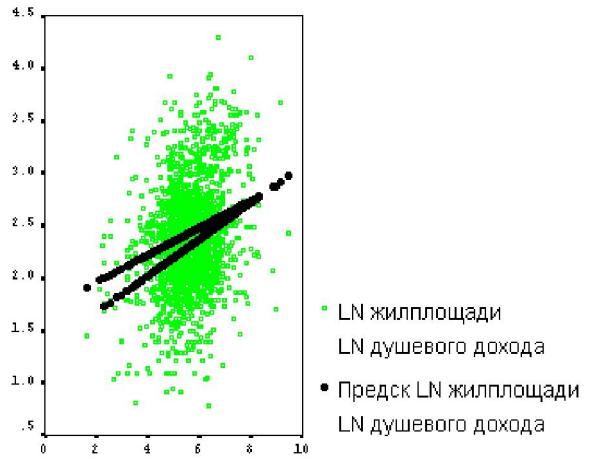

Рисунок 2. Линия регрессии для тестируемой группы ("курильшиков") и для ее дополнения

Стартовая позиция для "курильщиков" находится ниже (константа меньше), но коэффициент при независимой переменной побольше, поэтому линия регрессии тестируемой группы "догоняет" линию регрессии дополнения (см. рисунок 2).

Гипотезы

Попробуем содержательно обобщить полученные результаты и высказать гипотезы, которые могли бы быть проверены при дальнейшем изучением материала.

Группа покупателей табачных изделий имеет более стесненные жилищные условия, но больший разброс в жилплощади. Она имеет большие доходы и эти доходы теснее связаны с размером жилья. Можно предположить, что здесь сосредоточены семьи, имеющие относительно молодых мужчин, способных заработать "на табачок", но которые пока не имеют жилплощадь в том же объеме, что и остальные семьи. Возможно, среди некурящих достаточно много пожилых людей, вдов, давно получивших свое жилье, поэтому и связь с доходами у них слабее.

Заключение

Разработанный метод является пока лишь локальной победой над проблемой множественных сравнений. Гипотеза о независимости моделей и групп может уточняться. Возможно использование иных схем сравнений, например на основе совместных доверительных интервалов.

Тем не менее, как нам кажется, метод полезен с практической точки зрения, поскольку повышает надежность результатов. Метод не ограничен в развитии, поскольку позволяет подключать к нему множество моделей сравнения групп объектов.

Программное обеспечение полезно даже без множественных сравнений, поскольку предоставляет возможность в оперативном режиме формировать множество групп объектов и сравнивать их по множеству моделей.

Список литературы

1. Аптон Г. Анализ таблиц сопряженности. - М.: Финансы и статистика, 1982. -143с.
2. Аренс Х., Лейтер Ю. Многомерный дисперсионный анализ. - М.: Финансы и статистика, 1985.

3. Клейнен Дж. Статистические методы в имитационном моделировании. Выпуск 2 /М.: Статистика, 1978. стр. 169-217
4. Петрович М.Л., Давидович М.И. Статистическое оценивание и проверка гипотез на ЭВМ.- М.: Финансы и статистика, 1989. стр. 97-98
5. Российской мониторинг экономического положения и здоровья населения. Мир России. 1999. № 3
6. Ростовцев П.С. Статистическое согласование мер связи в анализе социально-экономической информации// М.: Экономика и математические методы. Том 26, 1991. стр. 149-156.
7. Ростовцев П.С. Статистические характеристики детерминации/ Статистическое моделирование экономических процессов. - Новосибирск: Наука, 1991.
8. Ростовцев П.С., Костин В.С., Корнюхин Ю.Г., Смирнова Н.Ю. Анализ структур социологических данных и их устойчивости. В монографии «Социальная траектория реформируемой России. Исследования новосибирской экономико-социологической школы. Новосибирск, Наука, 1999. с. 657-677.
9. Ростовцев П.С., Костин В.С., Олех А.Л. Множественные сравнения в детерминационном и типологическом анализе. // Анализ и моделирование экономических процессов переходного периода в России. Выпуск 3.- Новосибирск, ИЭиОПП СО РАН, 1998. с.209-222.
10. Ростовцев П.С., Костин В.С., Олех А.Л. Множественные сравнения в таблицах для неальтернативных вопросов// Анализ и моделирование экономических процессов переходного периода в России. Выпуск 4.- Новосибирск, ИЭиОПП СО РАН, 1999. с.148-164.
11. Хастингс Н., Пикок Дж. Справочник по статистическим распределениям – М.: Статистика, 1980.
12. Шеффе Г. Дисперсионный анализ. – М.: Наука, 1980.
13. Green W.H. Econometric analysis / New Jersey: Prentice Hall Inc., 1997, p337-352
14. Haberman Sh.J. Analysis of Qualitative Data. Vol. 1.New York: Academic press. 1978..
15. Lehmannn E.L. Nonparametrics: Statistical methods based on ranks. San Francisco: Holden-Day, 1975.
16. Lerman I.C. (1980) Combinatorial Analysis in Statistical treatment of Behavioral Data. Quality and Quantity, vol. 14, No.3, p 431-469.
17. SPSS 8.0 Base for Windows. Chicago, 1996.
18. Williams R. D.Multiple comparisons among correlation coefficients // The American Statistician, 1991. Vol.45, p.341- 341