ОШИБКИ ИНТЕРПРЕТАЦИИ РЕГРЕССИОННЫХ МОДЕЛЕЙ В ПСИХОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ*

В.М. КОЗУБОВСКИЙ, доктор психологических наук, профессор кафедры юридической психологии Минского института управления
Е.С. НЕЧАЕВА, сотрудник Института психологии Ягеллонского университета (Краков)

Дан анализ типовых ошибок, допускаемых психологами при смысловой трактовке результатов наблюдений, оформленных в виде зависимостей (моделей) регрессионного типа. В первую очередь это касается ошибочной интерпретации коэффициентов регрессии, которым неоправданно, вопреки базовым теоретическим положениям математической статистики, приписывают роль «весовых» и на этом основании используют регрессионные модели психологических объектов в качестве инструмента управления и оптимизации. Показано, что в большинстве случаев столь «насильственное» привлечение регрессионных моделей классического типа к решению несвойственных им задач приводит к внешне элегантным, но не отвечающим требованию адекватности результатам.

Описаны условия, при которых коэффициенты регрессии приобретают весовые свойства (проведение наблюдений в соответствии со специальной матрицей ортогонального типа). Указываются пути построения многофакторных регрессионных моделей психологических объектов, позволяющие избежать ошибочной интерпретации. Подчеркивается, что в вузах при изучении математических методов психологии необходимо обращать внимание студентов на вопросы корректного использования математического инструментария.

Статья рассчитана на читателя, знакомого с основами классического регрессионного анализа на уровне пользователя.

Ключевые слова: регрессионная зависимость, метод наименьших квадратов, ковариационная матрица, «вес» коэффициента регрессии, стратегии эксперимента, поле наблюдения, интерполяция, экстраполяция, управление, условие ортогональности.

Многофакторный характер психологических объектов требует при их исследовании от психолога изучения большого количества внутренних взаимосвязей (их степени коррелированности, причинно-следственных отношений, значимости влияния на выходные и промежуточные показатели, уровня нелинейности и др.). В большинстве диссертационных и научно-исследовательских работ по психологической проблематике изучение таких объектов ограничивается установлением корреляционных связей между исследуемыми факторами и попыткой дать этим связям психологическую интерпретацию. Конечно, при таком подходе можно рассчитывать только на поверхностное выявление *локальных*, фрагментарных свойств объекта. *Системные* же свойства, представляющие наибольший интерес для психолога, как правило, остаются за пределами его «видимости», скрытыми (как ни парадоксально это звучит, но это так) за большим массивом полученных данных. Возникает ситуация, которую физики называют «проклятьем размерности».

В своем стремлении к более целостному изучению объекта наиболее «продвинутые» в математическом инструментарии психологи обращаются к многомерным схемам организации наблюдений, базирующимся на классическом регрессионном анализе. При этом результаты наблюдений описываются в виде регрессионной модели. Если имеются основания считать допустимой гипотезу о *линейности* влияния исследуемых независимых факторов X на показатель Y, то используется модель вида

$$Y = a_0 + \sum_{i=1}^{k} a_i X_i = a_0 + a_1 X_1 + a_2 X_2 + ... + a_k X_k \quad (1)$$

Если гипотеза линейности неправомерна, то для описания выбираются нелинейные модели вида

$$Y = a_0 + \sum_{i=1}^{k} a_i X_i + \sum_{i,j=1}^{k} a_{ij} X_i X_j, \quad (2)$$

$$Y = a_0 + \sum_{i=1}^{k} a_i X_i + \sum_{i,j=1}^{k} a_{ij} X_i X_j + \sum_{i=1}^{k} a_{ii} X^2_i;\ i < j \quad (3)$$

В уравнениях (1), (2) и (3) приняты обозначения:

Y – показатель «поведения» объекта (наблюдаемая зависимая переменная);

* Научная идея публикации принадлежит В.М. Козубовскому, текстовая реализация – Е.С. Нечаевой.

X – факторы, воздействующие на психологический объект (наблюдаемые независимые переменные);

a – коэффициенты регрессионной модели, подлежащие определению;

i; j – индексы коэффициентов и факторов (i=0,1,2,...,к; j=0,1,2,...,к; i < j);

к – количество факторов, включенных в наблюдение;

$\sum$ – символ суммирования по всем i-м факторам (от 1 до «к»).

Процедура расчета коэффициентов регрессии «а», основанная на методе наименьших квадратов, обычно трудностей не вызывает и в современных условиях реализуется на компьютере с помощью типовых программ. Далее психологу предстоит «вдохнуть жизнь» в полученную регрессионную модель. Вот здесь при попытке увязать ее математические признаки с характеристиками исследуемого объекта и подстерегает опасность ошибочной интерпретации психологической сущности изучаемого объекта.

Причина искажения обычно заключается в неверной трактовке роли найденных величин и знаков коэффициентов регрессии. Этим коэффициентам, к сожалению, в психологических исследованиях сплошь и рядом приписывается роль «весов». Последнее проявляется в том, что *величина* коэффициента «а» отождествляется со степенью влияния фактора «X», перед которым он стоит, на показатель «Y» (т.е. считают, что чем больше величина коэффициента, тем сильнее проявляется это влияние). *Знаку* же коэффициента приписывается направление влияния этого фактора на показатель (в сторону увеличения величины показателя Y при a > 0 или уменьшения – при a < 0). Другими словами, полученная регрессионная модель неправомерно наделяется *управляющими* возможностями. Если, например, психологу требуется обеспечить максимальное значение показателя Y_{max}, то он, не задумываясь о последствиях, достигает этого путем увеличения значений факторов X, перед которыми стоят коэффициенты с наибольшими значениями и положительными знаками.

Бесспорно, это было бы очень удобно. Но, увы, желаемое здесь принимается за действительное. Дело в том, что коэффициенты регрессии, полученные на основе наблюдений согласно стратегии классического регрессионного анализа (типа «пришел – увидел – записал»), не обладают «весовыми» свойствами. Эта стратегия не в состоянии обеспечить автономность, независимость коэффициентов друг от друга по той простой причине, что «кусочки» одного коэффициента могут находиться в других. В этой связи эффект изменения величины показателя Y при изменении значения какого-то фактора принадлежит не только данному (варьируемому) фактору, но и другим факторам. Вот этот «групповой» феномен совместного влияния факторов при изменении одного из них на значения показателя Y не позволяет использовать математическую модель психологического объекта, построенную на базе классического регрессионного анализа, в качестве управляющей. По этой же причине недопустимо использовать ее как инструмент прямой оптимизации, имеющей целью поиск экстремального значения показателя Y за счет непосредственного и целенаправленного варьирования значений факторов X.

Психологами допускается еще один вид ошибок в процессе построения регрессионной модели. Так, определив факт статистической незначимости какого-то коэффициента регрессии по критерию Стьюдента, они без сомнений исключают фактор, перед которым он стоит, из полученной регрессии. Корректное же использование процедуры построения регрессии требует при этом обязательного пересчета всех других ее коэффициентов. И так необходимо поступать каждый раз, когда обнаруживается статистически незначимый коэффициент.

Какие же задачи может успешно решать психолог с помощью модели, полученной на базе стратегии наблюдения классического типа? Одна из них – задача *интерполяции*, «прогнозирования вовнутрь», с целью определения значений показателя Y, которые не были зафиксированы в процессе наблюдения. Но при этом необходимо следить, чтобы значения факторов X находились в пределах границ *поля наблюдения*, то есть в пределах тех границ, в которых проводились наблюдения на этапе построения регрессионной модели. Задачи интерполяции возникают перед психологом в тех случаях, когда по каким-то соображениям (безопасности, больших материальных или временных затрат) отсутствует возможность наблюдения за поведением показателя Y при желаемом конкретном наборе значений факторов X. Ведь у психолога отсутствует возможность по своему усмотрению устанавливать значения факторов X в каждом из N конкретных наблюдений.

Вторая задача – это задача *экстраполяции*, или «прогнозирования вовне поля наблюдения». Здесь психолог пытается найти значения

показателя Y при значениях факторов X, выходящих за интервалы, в которых они наблюдались на этапе построения регрессии. Однако здесь следует соблюдать осторожность: в полученную модель должны подставляться такие значения каждого из факторов, которые бы не превышали 15–20% от величины «своего» интервала.

При этом психолог должен быть уверен в том, что объект в новых условиях будет вести себя «без фокусов», то есть общая тенденция (закономерность) системного поведения исследуемого объекта сохранится прежней. В этих условиях (на слишком большом удалении значений факторов X от поля наблюдения) регрессионная модель не может предсказывать адекватного поведения объекта. Здесь можно провести аналогию с физическим явлением критической массы: стоит только превысить массу делящегося вещества, в которой может протекать самоподдерживающаяся цепная реакция деления атомных ядер, как резко нарушится закономерность реакции и последует взрыв.

Как же «научить» регрессионные модели (1), (2), (3) решать «престижные» задачи *управления*? Что является препятствием для их использования в этом качестве? Чтобы найти ответ на этот вопрос, необходимо обратиться к некоторым деталям организации наблюдений за психологическим объектом и отдельным процедурным моментам расчета коэффициентов регрессии.

Известно [6], что свойства регрессионной модели обусловлены свойствами ее *ковариационной матрицы* (4)

$$\begin{Vmatrix} s^2(a_0) & \mathrm{cov}(a_0\ a_1)\ldots\mathrm{cov}(a_0\ a_к) \\ \mathrm{cov}(a_1\ a_0) & s^2(a_1)\ldots\ldots\mathrm{cov}(a_1\ a_к) \\ \mathrm{cov}(a_2\ a_0) & \mathrm{cov}(a_2\ a_1)\ldots\mathrm{cov}(a_2\ a_к) \\ \ldots & \ldots \\ \mathrm{cov}(a_к\ a_0) & \mathrm{cov}(a_к\ a_1)\ldots\ldots s^2(a_0) \end{Vmatrix}$$

Иногда ее записывают в виде

$$s^2(Y)=\begin{Vmatrix} c_{00} & c_{01} & c_{0к} \\ c_{10} & c_{11} & c_{1к} \\ c_{20} & c_{21} & c_{2к} \\ \ldots & \ldots & \ldots \\ c_{к0} & c_{к1} & c_{кк} \end{Vmatrix},$$

где $s^2(Y)$ – дисперсия ошибки наблюдения за значениями показателя Y;

$c_{ii} = s^2(a_i) : s^2(Y)$; $c_{ij} = \mathrm{cov}(a_i\ a_j) : s^2(Y)$; $s^2(Y)$ – элементы матрицы;

$s^2(a_i)$ – диагональные элементы матрицы (4), характеризующие дисперсии коэффициентов регрессионной модели исследуемого объекта (ошибки в их определении);

$\mathrm{cov}(a_i\ a_j)$ – недиагональные элементы матриц (4), (5), характеризующие ковариации коэффициентов модели (статистическую зависимость между коэффициентами).

Элементы $s^2(a_i)$ и $\mathrm{cov}(a_i\ a_j)$ матрицы (4), (5) включены в формулы для расчета коэффициентов регрессии «а» [6]

$$a_i = \sum_{i=0}^{k} c_{ij}(i\,Y); \quad i = 0,1,2,\ldots,k; \ i < j, \qquad (6)$$

$$\text{где } (i\ Y) = \sum_{u=1}^{N} X_{iu} Y_u; \quad u = 1,2,\ldots,N;$$

N – общее число наблюдений;

X_{iu} – элементы матрицы [X] значений факторов в N наблюдениях;

$$[X] = \begin{Vmatrix} X_{01} & X_{11} & X_{к1} \\ X_{02} & X_{12} & X_{к2} \\ \ldots & \ldots & \ldots \\ X_{0N} & X_{1N} & X_{кN} \end{Vmatrix}$$

$$[iY] = \begin{Vmatrix} (0\ Y) \\ (1\ Y) \\ \ldots \\ (к\ Y) \end{Vmatrix} \qquad (8)$$

В матричной форме формула для определения коэффициентов в принятых выше обозначениях имеет вид

$$A = [X^TX]^{-1}\,X^TY, \qquad (9)$$

где X^T – транспонированная матрица [X].

Чтобы использовать регрессионную модель наблюдаемого объекта в целях *управления*, необходимо обеспечить независимость коэффициентов «а». А это значит, что матрица (4) должна быть *диагональной*, то есть все ее недиагональные элементы должны быть равны нулю [1]

$$\mathrm{cov}(a_i\ a_j) = 0 \qquad (10)$$

Только при таком условии появляется возможность автономного влияния каждого фактора на показатель Y с учетом знака коэффициента. Правомерным оказывается и ранжирование факторов по величине их коэффициентов. Наконец, психологу предоставляется возможность оптимизации модели объекта, т.е. поиска такой совокупности значений факторов $X_{орт}$, при которых достигается экстремальное значение показателя $Y_{эстр}$. Для решения задач оптимизации могут использоваться формальные методы исследования операций и математического анализа.

Кстати, диагональность матрицы (4) снимает требование обязательного пересчета коэффициентов регрессии после признания одного из них незначимым. Фактор при этом коэффициенте «безболезненно» исключается из модели.

Теперь наступил момент ответить на вопрос: как же обеспечить диагональность ковариационной матрицы (4) и тем самым получить все желаемые преимущества по использованию моделей регрессионного типа?

Из формул (6), (9) видно, что расчетные значения коэффициентов регрессии «a» обусловлены (помимо сказанного выше) еще и структурой матрицы [X]. Доказано, что если структура матрицы [X] отвечает так называемому условию *ортогональности*, при котором сумма почленных произведений любых двух столбцов матрицы [X] равна нулю, то есть

$$\sum_{u=1}^{N} X_{iu} Y_u = 0; \quad i = 0,1,2,..., k; \; i < j, \; u = 1,2,..., N, \quad (11)$$

то матрица $[X^T X]^{-1}$ оказывается диагональной

$$\begin{Vmatrix} 1/N & 0 & 0 \ldots\ldots 0 \\ 0 & 1/N & 0 \ldots\ldots 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ 0 & 0 & 0 \ldots 1/N \end{Vmatrix}$$

и формула (9) превращается в скалярное произведение столбца Y на соответствующий столбец Xi , то есть

$$a_i = \sum_{u=1}^{N} X_{iu} Y_u = 0; \quad i = 0,1,2,..., k; \quad i < j; \; u = 1,2,..., N \quad (12)$$

Все это свидетельствует о том, что коэффициенты регрессии приобретают свойство «весов» для факторов, перед которыми они стоят. Условие ортогональности (12) реализуется построением специальных матриц [X] до проведения наблюдений. Примером простейших матриц такого типа являются матрицы полного факторного эксперимента (ПФЭ). Более подробно с ними можно познакомиться, например, в [2; 3; 4; 5; 6].

Авторам представляется, что в учебных дисциплинах высших учебных заведений, касающихся математических методов в психологии, изложенные выше соображения должны находить детальное отражение. Это будет гарантировать более корректное использование аппарата регрессионного анализа в задачах экстраполяции, интерполяции и оптимизации.

ЛИТЕРАТУРА

1. Зедгендзе И.Г. Планирование эксперимента для исследования многокомпонентных систем. М.: Наука, 1976.
2. Козубовский В.М. Оптимальные стратегии поисковых и формирующих экспериментов в психологических исследованиях // Белорусский психологический журнал. 2004. № 1. С. 52–59.
3. Козубовский В.М. Моделирование в практической психологии: Учеб. пособие. Мн., РИПО, 1995.
4. Налимов В.В. Теория эксперимента. М.: Наука, 1971.
5. Планирование эксперимента в исследовании технологических процессов / Под ред. Э.К. Лецкого. М.: Мир, 1977.
6. Суходольский Г.В. Основы математической статистики для психологов. СПб., 1998.