УДК 519.23

Новый подход к оцениванию силы статистической связи регрессоров

Дронов С.В.
Алтайский государственный университет, г. Барнаул
dsv@math.asu.ru

Аннотация

В работе предложен новый коэффициент, позволяющий оценить ту составляющую статистической связи факторов множественной регрессии, которая отражает именно степень их взаимодействия при формировании выхода регрессии, отсекая части их связи, не оказывающие непосредственного влияния на выход. Даются рекомендации по применению этого коэффициента для повышения адекватности нелинейных регрессионных уравнений.

Ключевые слова: множественная регрессия, статистическая связь факторов регрессии, нелинейные регрессионные модели.

1. Регрессионная связь факторов над данным выходом.

Рассмотрим стандартную задачу регрессии: имеется зависимый показатель (выход, отклик) Y и несколько факторов $X_1, \ldots, X_q$ (регрессоров), вид и сила зависимости выхода от которых исследуется. Если используется линейная модель, и факторы в ней независимы или, по крайней мере, не коррелированы, то оценки всех параметров регрессии оказываются в определенном смысле наилучшими, а сама регрессия в приложениях дает наиболее точные результаты (см., например, [1, с. 105-108]). В общем случае связи между факторами в первом приближении учитываются в модели через величины внедиагональных элементов ковариационной матрицы исходных данных. За счет их появления матрица плана становится хуже обусловленной, что приводит как к вычислительным сложностям, так и меньшей устойчивости уравнения регрессии к возможным ошибкам выборочных данных и, как следствие, к более низкой адекватности этого уравнения, на что прямо указано в [2].

Но насколько хорошо элементы ковариационной матрицы отражают истинную картину связи между факторами регрессии? Известно, что такая матрица адекватно отражает величину линейной составляющей соответствующей связи, что, вероятно, достаточно при изучении линейной модели. Если же речь идет о силе связи в ее самом общем понимании, то, по-видимому, следует считать факторы в модели регрессии не влияющими друг на друга, только если функция регрессии связана с ними только “в чистом виде”, т.е. не включает в себя никаких их взаимодействий. Например, так бывает, когда наилучший прогноз величины выхода Y по значениям факторов имеет вид

$$\widehat{Y} = \phi(X_1, ..., X_q) = \sum_{i=1}^{q} \phi_i(X_i), \tag{1}$$

где функции ϕ_i, $i = 1, ..., q$ зависят только от своих аргументов (аддитивный случай). Возможны и другие варианты понимания тезиса об отсутствии взаимодействий факторов в функции регрессии, например, мультипликативный случай, когда функция представляется не суммой, а произведением множителей вида $\phi_i(X_i)$. Но, по мнению автора, (1) следует иметь в виду в первую очередь.

Если вдруг мы убедимся в том, что факторы-регрессоры действительно не влияют друг на друга в интересующем нас смысле, то это не является основанием утверждать, что они не связаны между собой вовсе. Интересующий нас вид связи – это связь факторов через посредство выхода, связь через схожесть их влияния на выход, а не собственно влияние одного из факторов на другой. "Полная" связь факторов, понимаемая в традиционном смысле, безусловно, включает в себя эту составляющую, но ее доля в общей силе связи может быть настолько малой, что, даже при очевидной их зависимости, вполне можно считать, что связь подобного типа между ними отсутствует.

Попытавшись представить сказанное наглядно, обратимся к известной геометрической интерпретации корреляционной связи. При представлении выборочных значений каждой из наблюдаемых переменных вектором соответствующей объему выборки размерности, степень связанности двух переменных определяется величиной проекции одного из таких векторов на направление другого. Подобная интерпретация проистекает из того, что, после центрирования координат каждого из векторов, выборочный коэффициент корреляции между ними оказывается равным косинусу соответствующего угла. После возврата к исходным (нецентрированным) векторам угол между ними меняется, но меньший угол будет переходить в меньший, а значит, большему коэффициенту корреляции по-прежнему будет соответствовать большая величина проекции одного вектора на другой.

Итак, именно коэффициентом корреляции проекций вектора-выхода на направления векторов-факторов и характеризуется степень интересующей нас связи факторов в модели линейной регрессии. Таким образом, в линейном случае степень связи двух факторов между собой определяется величиной косинуса угла между проекциями вектора выхода на векторы-факторы, – чем больше этот косинус (и, соответственно, меньше угол), тем сильнее связь. Разумеется, здесь угол между проекциями совпадает с углом между векторами-факторами.

Откажемся от линейности регрессии. Пусть, например, нам известна функция регрессии ϕ_i, обеспечивающая наилучший прогноз Y по значениям i-го фактора без учета остальных. Если критерием силы связи по-прежнему служит коэффициент корреляции, то вектор $Z_i = \phi_i(X_i)$ обладает тем свойством, что среди всех $\phi(X_i)$ проекция вектора выхода на него будет наибольшей по величине. Можно, разумеется, без ограничения общности считать, что величина этой проекции равна $|\phi_i(X_i)|$. Если рассмотреть два фактора X и Z, каждый со своей функцией регрессии, то величина коэффициента корреляции между преобразованными факторами и будет количественной оценкой силы той опосредованной связи, которой обладают факторы по отношению к показателю-выходу. Дадим определение.

Пусть задана некоторая переменная Y, которую мы назовем выходом. Коэффициентом регрессионной связи между переменными X и Z над выходом Y назовем величину

$$R^Y(X,Z) \;=\; |\rho(\phi_X(X), \phi_Z(Z))|, \tag{2}$$

где $\rho(\cdot,\cdot)$ – коэффициент корреляции (Пирсона), ϕ_X, ϕ_Z – функции регрессии Y на X и Z соответственно.

Мы не будем выходить в область отрицательных значений коэффициента, поскольку интересуемся силой связи, а не ее направлением. Введенный коэффициент обладает диапазоном значений $[0,1]$, причем нулевое его значение означает только отсутствие взаимодействия факторов в их влиянии на выход Y, а не отсутствие их связи между собой.

Частный случай подобного вида связи, когда набор некоторых объектов заданный своими показателями X, Z разбит на кластеры, а Y для каждого из объектов представляет собой числовую метку кластера, в который этот объект попал, изучался в [3]. Показатель Y в этом случае называется кластерной переменной. Тот же вид связи, который при этом выделяется, был в цитированной работе назван пост-кластерной связью.

Отметим еще раз вслед за [3], что, введя новый коэффициент корреляции (2) между показателями X и Z, мы фактически определили новый вид статистической связи между ними. Назовем этот вид регрессионной связью над выходом Y.

2. Оценивание коэффициента регрессионной связи.

Расчет введенного выше коэффициента регрессионной связи на практике связан с проблемой оценивания функции регрессии. Для подобных оценок разработано много методов, см., например [4, 5]. Вероятно, самый простой способ выглядит так. Пусть мы оцениваем коэффициент (2) по результатам n наблюдений вида $(x_j; z_j|y_j)$, $j = 1, ..., n$, X, Z, Y – n-мерные векторы с соответствующими координатами (выборки). Зафиксируем класс допустимых преобразований факторов Φ. Условимся считать, что, если $X = (x_1, ..., x_n)$, $\phi \in \Phi$, то $\phi(X) = (\phi(x_1), ..., \phi(x_n))$. Для каждого из показателей X и Z в отдельности найдем такое преобразование из класса Φ, что его результат наилучшим образом коррелирует с выходом:

$$\phi_X^* = \arg \max_{\phi\in\Phi} |\rho(\phi(X), Y)|, \quad \phi_Z^* = \arg \max_{\phi\in\Phi} |\rho(\phi(Z), Y)|.$$

Здесь ρ – выборочный коэффициент корреляции Пирсона. Выборочный коэффициент регрессионной связи над выходом Y находим, используя (2):

$$R^{Y*}(X, Z; \Phi) = |\rho(\phi_X^*(X), \phi_Z^*(Z))|. \tag{3}$$

Предложение 1. *Пусть Φ_k – класс многочленов степени не выше заданного $k \in N$. Тогда справедливы следующие утверждения*

1. *$R^{Y*}(X, Z; \Phi_1)$ совпадает с $|\rho(X, Z)|$.*
2. *Если ни в X, ни в Z нет одинаковых элементов, то $R^{Y*}(X, Z; \Phi_{n-1}) = 1$.*

Доказательство. Первое очевидно из определения (3) и того известного факта, что модуль коэффициента корреляции инвариантен относительно линейных невырожденных преобразований. Второе следует из того, что для произвольных n точек с различными абсциссами всегда существует многочлен степени не выше $n - 1$, график которого проходит через все эти точки. □

Предложение 2. *При выборе оценок функций регрессии из класса Φ_k величина $|\rho(\phi_X^*(X), Y)|$ не убывает при увеличении k. Если в X нет одинаковых элементов, то $\rho(\phi_X^*(X), Y) = 1$ при $k \geqslant n - 1$.*

Доказательство. Неубывание $|\rho|$ следует из геометрической интерпретации коэффициента регрессионной связи: более широкий класс преобразованных векторов-факторов позволяет выбрать в нем вектор имеющий, по крайней мере, не меньшую проекцию вектора-выхода на него, чем на имевшийся ранее. Далее, если график многочлена ϕ_X^* проходит через все точки (x_i, y_i), $i = 1, ..., n$, то ясно, что $\rho(\phi_X^*(X), Y) = \rho(Y, Y) = 1$. □

3. О построении примеров слабой регрессионной связи.

После анализа определения (3) может возникнуть впечатление, что, поскольку преобразованные факторы по отдельности сильно коррелируют с выходом Y, то коэффициент регрессионной связи всегда будет достаточно велик. Это не вполне справедливо, но произвольно малым он, вообще говоря, все же быть не может. Точную нижнюю его границу дает следующее утверждение.

Предложение 3. *Пусть имеются три показателя с конечными ненулевыми дисперсиями. Тогда для выборочных коэффициентов корреляции, вычисленных по трем связанным выборкам X, Y, Z справедливо неравенство*

$$|\rho(X,Z)| \geqslant |\rho(X,Y)\rho(Z,Y)| - \sqrt{(1-\rho^2(X,Y))(1-\rho^2(Z,Y))},$$

причем равенство достигается, когда $\rho(X,Y), \rho(Z,Y) \geqslant \sqrt{2}/2$, а выборки X, Y, Z как многомерные векторы располагаются в одной плоскости, т.е. линейно зависимы.

Доказательство. Центрируем все выборки и отложим получившиеся векторы от начала координат в евклидовом пространстве соответствующей размерности. Тогда косинусы плоских углов образовавшегося трехгранного угла равны рассматриваемым коэффициентам корреляции. Известна (например, см. [6, с. 205]) следующая трехмерная теорема косинусов:

$$\cos\alpha = \cos\beta\cos\gamma + \sin\beta\sin\gamma\cos A, \tag{4}$$

где α, β, γ плоские углы, а двугранный угол A лежит напротив α. Переходя в этой формуле к коэффициентам корреляции, приходим к неравенству

$$|\rho(X,Z)| \geqslant |\rho(X,Y)\rho(Z,Y)| - \sqrt{(1-\rho^2(X,Y))(1-\rho^2(Z,Y))}\cdot \cos A,$$

что и доказывает неравенство предложения, если мы оценим $\cos A$ сверху единицей. Для того, чтобы выяснить условия, при которых происходит достижение равенства, рассмотрим элементарное неравенство, применением которого к (4) получается требуемая оценка:

$$|a+b| \geqslant ||a|-|b|| \geqslant |a|-|b|.$$

Равенство в нем достигается тогда и только тогда, когда выполнены два условия: $|a| \geqslant |b|$ и a, b разных знаков. Первое условие выполняется, например, когда каждый из косинусов больше соответствующего синуса, что справедливо при $\rho(X,Y), \rho(Z,Y) \geqslant \sqrt{2}/2$. Поскольку оба синуса в (4) неотрицательны, то для выполнения второго условия угол A должен быть тупым. Наименьшее же значение получается для случая развернутого двугранного угла A, т.е. плоской картинки. Предложение доказано. □

Предложение 4. *Имеет место оценка*

$$\left|R^{Y*}(X,Z;\Phi) - \rho(\phi_X^*(X),Y)\rho(\phi_Z^*(Z),Y)\right| \leqslant \sqrt{(1-\rho^2(\phi_X^*(X),Y))(1-\rho^2(\phi_Z^*(Z),Y))}.$$

Это неравенство вытекает из представления соответствующих коэффициентов корреляции через косинусы плоских углов трехгранного угла и равенства (4). Заметим, что если оценки функции регрессии дают большие по модулю коэффициенты корреляции преобразованных факторов с выходом, то доказанное можно интерпретировать как

$$R^{Y*}(X,Z;\Phi) \approx \rho(\phi_X^*(X),Y)\rho(\phi_Z^*(Z),Y).$$

Из предложения 3 следует, что для построения примера показателей, не коррелированных между собой, но относительно сильно коррелированных с выходом Y, необходимо разместить представляющие их векторы в одной плоскости с вектором выхода. При этом углы между каждым из них и вектором-выходом должны в сумме составлять прямой угол. Поскольку степень коррелированности тем больше, чем меньше соответствующий угол, то для каждого из фактором его углы с выходом имеет смысл взять равными по величине. Итак, один из векторов-факторов должен быть симметричен другому относительно вектора-выхода.

Приведем здесь простой способ построения вектора Z, симметричного заданному вектору X относительно третьего вектора Y и лежащего в той же двумерной плоскости, что и векторы X, Y. Пусть

$$P = \frac{|X|\cos(\widehat{X,Y})}{|Y|} \cdot Y -$$

проекция X на Y . Тогда по правилу параллелограмма достаточно взять

$$Z = 2P - X. \tag{5}$$

Обратимся к числовому примеру. Возьмем $Y = (1;\ 1;\ 2;\ 3\,;3)$. Среднее значение координат вектора Y равно 2. Центрируя его, получим $(-1; -1; 0; 1; 1)$. Вектор $(\sqrt{2}-1;\ -\sqrt{2}-1;\ 0; 1; 1)$ расположен к Y под углом $\pi/4$. Применение (5) дает симметричный вектор $(-1-\sqrt{2};\ \sqrt{2}-1;\ 0; 1; 1)$. Возвращаясь к исходному началу координат, видим, что, выбирая

$$X = (1+\sqrt{2};\ -\sqrt{2}+1;\ 2; 3; 3), \quad Z = (1-\sqrt{2};\ \sqrt{2}+1;\ 2; 3; 3),$$

придем к значениям $\rho(X,Y) = \rho(Z,Y) = \sqrt{2}/2 \approx 0,707, \quad \rho(X,Z) = 0$. Итак, при высокой степени коррелированности каждого из X, Z с Y корреляция между X и Z отсутствует.

Приведенный пример формально может полностью нас устроить только в случае, когда мы работаем с линейной регрессией. Действительно, при расчете коэффициента регрессионной связи мы имеем дело не с самими векторами факторов, а с преобразованными их вариантами. И, если в примере можно было X считать уже преобразованным вариантом исходных данных, то всегда ли будет Z, полученный симметрией X относительно Y, результатом некоторого оптимального преобразования? Как правило, ответ на этот вопрос положителен, но выяснение необходимых и достаточных условий этого и подробное их обоснование требует отдельного исследования.

Вместо этого ограничимся здесь парой примеров. Рассмотрим Φ_2 в качестве класса допустимых преобразований (квадратичная регрессия), и пусть

$$Y = (1;\ 1;\ 2;\ 3;\ 3),\ X = (-4;\ 4;\ 2;\ 3;\ 3),\ Z = (0;\ 5;\ -1;\ 3;\ 0).$$

Тогда

$$\phi^*_X(x) = -0,11x^2 + 0,08x + 3,1, \quad \phi^*_Z(x) = -0,16x^2 - 0,57x + 2,32,$$

$$\phi^*_X\left(\overrightarrow{X}\right) = (0,97;\ 1,60;\ 2,80;\ 2,31;\ 2,31), \quad \phi^*_Z\left(\overrightarrow{Z}\right) = (2,32;\ 1,16;\ 1,59;\ 2,59;\ 2,33),$$

а коэффициент регрессионной связи (3) над выходом Y равен –0,02, хотя до преобразований это значение составляло $\rho(X,Z) = 0,49$.

Если вместо этого, оставив X, Y теми же, но выбрать $Z = (2;\ 2;\ -9;\ 3;\ 3)$, то несложные вычисления дают

$$\rho(X,Z) = -0,02; \quad R^{Y*}(X,Z;\Phi_2) = 0,717,$$

что доставляет пример противоположного случая.

4. Обсуждение и выводы.

Первое утверждение предложения 1 показывает, что рассматриваемый нами новый вид связи для случая линейной регрессионной модели совпадает с обычной корреляционной связью. Это можно также интерпретировать, как совпадение выделяемой регрессионной связи над выходом Y со всей имеющейся связью в линейном случае. Следовательно, введенное понятие может выявлять ранее не учтенные эффекты только в нелинейных регрессионных моделях. Предложение 2 дает еще один аргумент в пользу того, что рассмотрение

полиномиальных регрессий высоких степеней лишено смысла. Отметим по ходу дела, что использование многочленов степени лишь не выше третьей в качестве возможной оценки функции регрессии – обычная практика.

Второе утверждение предложения 1 показывает бесперспективность совместного использования в модели регрессии сильно коррелированных факторов: при попытке оценить, так сказать, "нелинейную составляющую" их связи, никаких новых выводов не последует. Действительно, коэффициент регрессионной связи будет принимать только значения, большие абсолютной величины коэффициента корреляции Пирсона, т.е., вне зависимости от выбора степени аппроксимирующего многочлена, связь факторов будет интерпретироваться как сильная, что было ясно и в линейной модели.

Основное применение нового вида статистической связи, рассмотренного в работе – изучение двух слабо коррелированных факторов в нелинейной регрессионной модели. Традиционные методики в таком случае предлагают дальнейшее использование обоих факторов. Как мы уже видели, это оправдано при рассмотрении линейной регрессии. Если же нас интересуют более общие ситуации, то здесь возможно, что, хотя факторы и практически не коррелированы, регрессионная связь их над изучаемым выходом окажется довольно сильной.

Например, если для X, Z из рассмотренного выше первого примера подобрать оптимальные преобразования класса Φ_2 (оценка функции регрессии подходящим многочленом не выше второй степени), то получим

$$\phi^*_{\overrightarrow{Z}}(x) = 0,38x^2 - 0,46x + 0,77 = \phi^*_Z(x), \quad R^{Y*}(X, Z; \Phi_2) \approx 0,726\,,$$

хотя выборочная корреляция между X и Z отсутствовала.

Таким образом, можно попытаться заменить в модели два изучаемых фактора на один вопреки классическим рекомендациям, если нас интересуют квадратичные зависимости или зависимости более высоких степеней. В итоге мы получим более простое регрессионное уравнение и сократим, например, размерность данных, необходимых для прогноза выхода Y.

Список литературы

1. Дронов С.В. Методы и задачи многомерной статистики. — Барнаул : Изд-во Алт. ун-та, 2015. — 275 с.

2. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. Множественная регрессия. — 3-е изд. — М. : Диалектика, 2016. — 912 с.

3. Дронов С.В., Леонгардт К.А. Оценивание силы пост-кластерной связи между формирующими показателями // МАК: "Математики - Алтайскому краю": сборник трудов всероссийской конференции по математике. — Барнаул : Изд-во Алт. ун-та, 2018. — С. 26–29.

4. Браништи В.В. Оптимизация алгоритмов настройки коэффициента размытости для непараметрических оценок // Молодежь и наука: сборник материалов всероссийской научно-технической конференции. — Красноярск : Сиб. федер. ун-т, 2014. — URL: http://conf.sfu-kras.ru/sites/mn2014/pdf/d02/s14/s14{_}002.pdf.

5. Лапко А.В., Лапко В.А. Непараметрические модели и алгоритмы обработки информации. — Красноярск : Изд-во Сиб. гос. аэрокосмич. ун-та, 2010. — 220 с.

6. Богуш А.А. Избранные труды. — Минск : Беларус. навука, 2011. — 578 с.