УДК 519.233.5, 51.76

10.56564/27825264_2026_1_50

О ПОНЯТИИ «КОРРЕКТИРОВКА НА КОНФАУНДЕРЫ» В ЛИНЕЙНЫХ РЕГРЕССИОННЫХ МОДЕЛЯХ

А. Н. Вараксин

Институт промышленной экологии УрО РАН, г. Екатеринбург, Россия

Методологическая статья посвящена проблемам, возникающим при трактовке результатов множественной линейной регрессии (МЛР) с коррелированными предикторами. Корреляция между предикторами X приводит к тому, что коэффициенты множественной регрессии, описывающие воздействие на отклик Y некоторого комплекса предикторов, оказываются разными в зависимости от набора предикторов, входящих в модель МЛР. В частности, они отличаются от коэффициентов моделей простой регрессии, каждая из которых описывает влияние на отклик одного из предикторов. Коэффициенты простой регрессии принято называть нескорректированными (crude), а коэффициенты МЛР – скорректированными (adjusted) на остальные предикторы, входящие в МЛР. Такой подход называется «корректировка на конфаундеры». В данной работе показано, что эффект влияния какого-либо предиктора X_i на отклик Y описывается не коэффициентом регрессии b_i для этого предиктора, а всей функцией регрессии $Y=Y(x_1, x_2, ..., x_k)$, зависящей от всех предикторов $x_1, x_2, ..., x_k$. Анализ этого выражения позволяет утверждать, что влияние предиктора X_i на отклик Y в конечном счете описывается однофакторной функцией $Y=Y(x_i)$, построенной на основе расчетных Y_{calc} $(x_1, x_2, ..., x_k)$ или наблюдаемых Y_{obs} значений отклика Y и значений данного предиктора x_i. Показано, что процедура «корректировка на конфаундеры» не имеет математического обоснования и не может использоваться при анализе моделей множественной линейной регрессии.

Ключевые слова: множественная линейная регрессия; аддитивные и неаддитивные модели; коррелированные предикторы; корректировка на конфаундеры; однофакторные функции отклика.

1. Введение

В математической статистике линейными называют регрессионные модели, которые линейны по *коэффициентам* (коэффициенты b_i входят в модель в первой степени). Эти модели можно разделить на линейные и нелинейные по *предикторам*. Линейные по предикторам x_i модели множественной линейной регрессии (МЛР) имеют вид [1, 2]:

$$Y(x_1, x_2, x_3, ..., x_k) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + ... + b_kx_k, \quad (1)$$

где Y – отклик на воздействие предикторов X (x_1, x_2, …, x_k), b_i – коэффициенты модели. В линейных по предикторам моделях (1) влияния различных предикторов x_i на отклик Y просто суммируют, поэтому их можно назвать *аддитивными* моделями [1–3].

Наиболее распространенные и простые модели, нелинейные по предикторам x_i, содержат не только линейные по x_i члены, но и члены более высокого порядка, например, степени x_i, начиная со второй, либо так называемые перекрестные члены типа $(x_i \cdot x_j)$, в которых значения предикторов x_i и x_j перемножаются [1–3]. В

нелинейных по предикторам моделях нет простого суммирования эффектов отдельных предикторов, поэтому такие модели называют *неаддитивными* [1, 2, 4].

2. Регрессионные модели, линейные по предикторам

Допустим, имеется фактический (биологический, эпидемиологический) материал, содержащий численные значения отклика *Y* и значения *k* предикторов x_i (*i*=1, 2, 3, ..., *k*) для *n*-индивидуумов (*j*=1, 2, 3, ..., *n*). Используя этот материал, можно построить *k* моделей простой линейной регрессии (ПЛР), каждая из которых содержит один предиктор из числа имеющихся в наличии x_i

$$Y = a_{0(i)} + a_i x_i, (i = 1,2,3,..., k). \quad (2)$$

Кроме однофакторных моделей типа (2), можно построить с использованием того же фактического материала различные модели множественной регрессии (МЛР), содержащие два, три, ..., *k* предикторов; одна из таких моделей, содержащая все *k* предикторов, представлена формулой (1).

Если предикторы x_i не коррелируют друг с другом, то тогда коэффициенты b_i модели (1) с *k* предикторами совпадают с *k* коэффициентами a_i (*i*=1, 2, …, *k*) моделей (2). При наличии корреляций между предикторами коэффициенты b_i отличаются от коэффициентов a_i.

Особенностью аддитивных моделей (1) является то, что влияние каждого (одного) предиктора x_i на отклик *Y* определяется одним слагаемым $b_i x_i$

$$Y(x_i) \sim b_i x_i, \quad (3)$$

а изменение отклика *Y* при изменении x_i на единицу определяется одним коэффициентом b_i

$$\Delta Y(\Delta x_i = 1) = b_i \quad (4)$$

и не зависит от значений других предикторов x_j, их коэффициентов b_j и от наличия корреляций между предикторами.

Между тем наличие корреляций между предикторами накладывает определенные ограничения на методику расчета $\Delta Y(\Delta x_i)$. Корреляционная связь между предикторами, например, между x_i и x_j, выражается в том, что при изменении значений одного предиктора x_i значения второго предиктора x_j в *среднем* изменяются синхронно. Если между предикторами x_i и x_j имеется, например, *положительная* корреляционная связь (это значит, что при увеличении x_i значения x_j в среднем увеличиваются), то тогда при увеличении x_i мы должны рассчитывать $\Delta Y(\Delta x_i)$ с учетом того, что значения x_j также увеличиваются. Соотношения (3), (4) в аддитивных моделях (1) не дают нам такой возможности.

Трактовка выражения $\Delta Y(\Delta x_i = 1) = b_i$ в эпидемиологии. Для линейных по предикторам регрессионных моделей в эпидемиологической практике принято считать, что коэффициенты b_i модели (1) описывают *воздействие* предиктора x_i на отклик *Y*, *скорректированное* на остальные предикторы x_j, входящие в модель (1). Эти остальные предикторы называют *конфаундерами*, в связи с чем используется термин «корректировка на конфаундеры» (adjustment for confounders) или «контроль конфаундинга» (control of confounding). Такой подход излагается, например, в монографии Basic Epidemiology [5], изданной Всемирной организацией здравоохранения. Сообщения о том, что эффект какого-то предиктора в регрессионной модели рассчитан с корректировкой на конфаундеры, можно встретить в многочисленных журнальных публикациях (см., например, [6]). При

таком подходе считается, что коэффициент регрессии b_i описывает *реальное* воздействие конкретного предиктора X_i на отклик Y при наличии других предикторов и может быть использован для предметной (биологической, эпидемиологической) трактовки действия каждого X_i на Y в духе (3), (4).

В противоречие с данной точкой зрения специалисты по математической статистике подчеркивают, что соотношения (3) и (4) справедливы только в том случае, когда при изменении предиктора x_i значения всех остальных предикторов x_j считаются неизменными [2, 3, 7, 8]. Например, Э. Сигел пишет: «Коэффициент регрессии для каждого X определяет влияние этого X на Y при условии, что все остальные X не меняются: коэффициент регрессии b_i для i-й X-переменной указывает, какое увеличение Y ожидается, когда все X-переменные остаются неизменными, за исключением переменной X_i, которая увеличивается на одну единицу» [2, с. 614].

Следовательно, соотношение (4) выполняется только тогда, когда один предиктор может изменяться независимо от другого, что возможно только при отсутствии корреляционных связей между предикторами. При отсутствии этих корреляционных связей коэффициент b_i множественной регрессии (1) совпадает с коэффициентом a_i простой регрессии (2): $b_i=a_i$ и тогда при построении МЛР никакой корректировки на конфаундеры не происходит [1, 2].

Сейчас нелегко установить, каким образом методика [5] «корректировки на конфаундеры» в стиле соотношений (3), (4) была внедрена в эпидемиологическую практику, поэтому обратимся к современным и значимым изданиям, например, к самой монографии [5]. В этой монографии речь идет о «контроле конфаундинга» на этапе *анализа данных* двумя методами [5, p. 56]:

1 – стандартизация путем стратификации;

2 – статистическое моделирование.

В качестве примера стандартизации в главе 2 монографии [5] (раздел: Age-standardized rates) рассматривается показатель *смертности*, стандартизованный по возрасту (age-standardized *death* rate). Для получения этого показателя выполняется процедура стандартизации (*standardization*) путем стратификации индивидуумов по возрасту и последующего уравнивания по возрасту показателей смертности стратифицированных групп индивидуумов. Стандартизация путем стратификации является хорошо известной процедурой, все операции которой строго прописаны. Необходимость именно статистического моделирования, по мнению авторов [5], возникает тогда, когда конфаундеров становится слишком много и для выполнения многомерной стратификации требуется очень большое число индивидуумов. Авторы [5] считают, что статистическое моделирование (например, регрессия) свободно от такого ограничения на число индивидуумов.

Авторы [5] явно не говорят, но неявно подразумевают, что многомерная стратификация и моделирование дают один и тот же результат, если эти два подхода применить к одним и тем же данным. Такой вывод можно сделать из следующих положений монографии [5]. Термин *adjustment* (корректировка) первый раз встречается в уже упоминавшейся главе 2 (раздел: Age-standardized rates). В этом же разделе термин *standardization* употребляется как эквивалент термину *adjustment* (в оригинале: age-standardized death rate also referred to as an age-adjusted rate [5]). После того как в главе 2 термин adjustment был *объявлен* эквивалентом термину standardization, термин adjustment без объяснений был перенесен на регрессионные модели, которые в [5] считаются одним из методов статистического моделирования контроля конфаундинга (глава 4 в [5]). Никаких доказательств эквивалентности методов стандартизации и регрессии не приводится. При этом не дается никаких объяснений, почему коэффициент b_i у предиктора x_i в модели МЛР (1) является результатом процедуры, аналогичной процедуре стандартизации. Теория регрессии также не

дает на этот счет никаких указаний: метод получения коэффициентов b_i (метод наименьших квадратов) ничем не напоминает процедуру стандартизации [8].

Специалисты в области математической статистики доказывают, что коэффициенты b_i в модели (1) с коррелированными предикторами не могут трактоваться как $b_i = \Delta Y(\Delta x_i = 1)$ и не имеют предметного (биологического, эпидемиологического) смысла [2, 3, 8]. В монографии [8], изданной около 50 лет назад, показано, что включение очередного предиктора x_j, который коррелирует с уже включенными в модель предикторами, изменяет все имеющиеся в (1) коэффициенты b_i таким способом, который не имеет ничего общего с процедурой стандартизации; при этом изменения коэффициентов b_i, конечно, описываются определенными математическими соотношениями (см. ниже раздел 2). Факт отсутствия
у коэффициентов b_i модели (1) именно *предметного смысла* описан в монографии [8] в главе 13 со звучащим названием «Беды регрессионных коэффициентов» (в оригинале: Woes of Regression Coefficients).

Еще в одной монографии [9] на конкретном примере показано, как изменяются коэффициенты b_i модели (1) при минимальных изменениях способов определения этих коэффициентов. Автор [9], как и авторы [8], утверждает, что искать какой-либо предметный смысл в коэффициентах b_i модели МЛР (1) с коррелированными предикторами бессмысленно. Аналогичной точки зрения придерживается автор [2].

Таким образом, даже в регрессионной модели, линейной по предикторам (в аддитивной модели), коэффициенты b_i не могут служить показателем воздействия каждого предиктора x_i на отклик Y. Налицо явное противоречие между профессиональными математиками, которые разрабатывают регрессионные методы анализа данных [1–3,7, 8], и эпидемиологами [5], которые используют эти методы для решения конкретных задач.

3. Математический смысл коэффициентов аддитивной модели множественной регрессии с коррелированными предикторами

В разделе 2 были приведены аргументы в пользу того, что коэффициенты b_i аддитивной модели множественной регрессии (1) с коррелированными предикторами не имеют *предметного* (биологического, эпидемиологического) смысла. В данном разделе в виде двух *утверждений* будет показан их *математический* смысл.

Утверждение 1

Выше обсуждалось выражение $\Delta Y(\Delta x_i = 1) = b_i$ (4), описывающее изменение отклика Y при изменении какого-либо предиктора x_i модели (1), в котором не учитываются корреляции между предикторами. Корреляции между предикторами приводят к тому, что изменение предиктора x_i должно сопровождаться изменением всех других предикторов x_j. Модель (1) в принципе не позволяет учесть такие изменения, поэтому для учета корреляций при расчете $\Delta Y(\Delta x_i = 1)$ в модель (1) необходимо внести какие-то предположения, которые не содержатся явно в самой модели (1).

Abbott и Carroll [10] предложили учитывать корреляции между предикторами в модели (1) следующим образом: пусть для простоты в модели типа (1) имеется два предиктора x_1 и x_2

$$y = b_0 + b_1 x_1 + b_2 x_2 . \quad (5)$$

Без учета корреляций между предикторами изменение отклика Y при изменении x_1 описывается приведенным выше соотношением (4)

$$\Delta Y(x_1, \Delta x_1) = b_1 \cdot \Delta x_1,$$

поскольку предиктор x_2 при таком подходе считается неизменным. При учете корреляций между x_1 и x_2 изменение x_1 должно сопровождаться изменением x_2 (изменением в среднем). Тогда, согласно [10], изменение Y при изменении x_1 должно выражаться соотношением

$$\Delta Y(x_1, \Delta x_1) = b_1 \cdot \Delta x_1 + b_2 \cdot \Delta x_2, \quad (6)$$

где Δx_2 – корреляционное изменение x_2 при изменении x_1. Для определения Δx_2 авторы [10] предложили использовать явный вид связи x_2 с x_1 в виде уравнения простой линейной регрессии

$$x_2 = c_{02} + c_{21} x_1, \quad (7)$$

откуда можно определять Δx_2 по формуле

$$\Delta x_2 = c_{21} \cdot \Delta x_1. \quad (8)$$

Такое изменение предиктора x_2 должно привести, согласно (6), к изменению отклика Y на величину

$$\Delta Y(\Delta x_2) = b_2 \cdot \Delta x_2 = b_2 \cdot c_{21} \cdot \Delta x_1.$$

Тогда *полное* изменение отклика Y (6) при изменении x_1 с учетом корреляционного изменения x_2 имеет вид

$$\Delta Y(\Delta x_1) = b_1 \cdot \Delta x_1 + b_2 \cdot \Delta x_2 = (b_1 + b_2 \cdot c_{21}) \cdot \Delta x_1. \quad (9)$$

Это и есть способ, которым авторы [10] предлагали рассчитывать изменение отклика Y с учетом корреляции между x_1 и x_2 в модели (5). В работе [10] методику (6)–(9) предлагалось использовать для моделей *логистической* регрессии. Если же использовать эту методику для *линейной* регрессии, получается неожиданный результат [11, 12]:

$$b_1 + b_2 \cdot c_{21} = a_1, \quad (10)$$

где a_1 – коэффициент модели *простой* линейной регрессии, связывающей отклик Y с одним предиктором x_1

$$Y = a_{01} + a_1 x_1. \quad (11)$$

Получается, что в модели множественной линейной регрессии (5) с *двумя* предикторами x_1 и x_2 изменение отклика Y при изменении предиктора x_1 с учетом *линейного* корреляционного изменения x_2 оказывается равным таковому в модели простой регрессии (11) с *одним* предиктором x_1

$$\Delta Y(\Delta x_1) = a_1 \cdot \Delta x_1. \quad (12)$$

Соотношения (10) и (12), сформулированные выше для двух предикторов, оказываются справедливыми для любого числа предикторов. Так, для модели (1), включающей k предикторов, имеем

$$b_1 + b_2 \cdot c_{21} + b_3 \cdot c_{31} + ... + b_k \cdot c_{k1} = a_1 .$$

Аналогичное соотношение выполняется в модели (1) для любого предиктора x_i [13]. Окончательно, при выполнении условий типа (8) и (9) для модели (1) с коррелированными предикторами,

$$\Delta Y(\Delta x_i = 1) = a_i, \quad (13)$$

что кардинально отличается от (4).

Утверждение, аналогичное (13) для аддитивной модели (1) с коррелированными предикторами, предлагает Э. Сигел [2, с. 621], правда, без доказательств.

Для случая двух переменных x_1 и x_2 из равенства (10) для коэффициента a_1 (и аналогичного равенства для a_2) можно получить выражение для коэффициента b_1 модели (5) множественной регрессии в виде

$$b_1 = \frac{a_1 - a_2 \cdot c_{21}}{1 - r_{12}^2} . \quad (14)$$

В формулу (14) входят характеристики a_1 и a_2 однофакторных моделей типа (11), а также показатель c_{21} статистической связи (6) между самими предикторами, r_{12} – коэффициент корреляции между x_1 и x_2, а $c_{12} \cdot c_{21} = r_{12}^2$.

Аналогично получается выражение для коэффициента b_2

$$b_2 = \frac{a_2 - a_1 \cdot c_{12}}{1 - r_{12}^2} . \quad (15)$$

Формулы (14), (15) были представлены также в работах [7] и [9]. Эти формулы показывают, как коэффициенты b_i МЛР (5) связаны с коэффициентами a_i моделей простой регрессии. Напомним, что коэффициенты b_i ранее назывались *adjusted*, а коэффициенты a_i – *crude*. Согласно (14), коэффициент b_1 предиктора x_1, скорректированный (*adjusted*) на x_2, равен коэффициенту a_1 (*crude*) того же предиктора x_1 в модели простой регрессии (11), скорректированному на величину $(a_2 \cdot c_{21})$. Таким образом, формулы (14), (15) раскрывают математический смысл эпидемиологического понятия «корректировка на конфаундеры». Очевидно, что эта корректировка не имеет ничего общего с процедурой стандартизации и предметный (эпидемиологический) смысл коэффициентов b_i остается неопределенным.

Утверждение 2

Для аддитивной модели МЛР (1) с коррелированными предикторами Мостеллер и Тьюки [8] провели анализ математического смысла коэффициентов b_i. Авторы [8] показали, что при отсутствии предметного (эпидемиологического) смысла коэффициентов b_i эти коэффициенты имеют ясный математический смысл. Мостеллер и Тьюки пишут [8, с. 43]: «…коэффициент любого носителя, полученный методом наименьших квадратов, можно рассматривать как коэффициент регрессии отклика на этот носитель, скорректированный линейно по остальным носителям из его подгенератора».

Может показаться, что это утверждение о *корректировке* коэффициента b_i предиктора x_i по остальным предикторам x_j аналогично эпидемиологическому утверждению [5], однако это не так. В эпидемиологическом подходе авторы [5] пишут о коэффициенте b_i как о скорректированном показателе связи отклика Y с предиктором x_i без объяснения термина «корректировка». У Мостеллера и Тьюки [8]

речь идет о конкретной и четко описанной процедуре.

Мостеллер и Тьюки [8] описывают следующий вариант линейной корректировки:

– пусть имеется модель МЛР (5) с двумя предикторами x_1 и x_2;

– пусть также имеется модель простой регрессии с одним предиктором x_2 вида, аналогичного модели (11).

$$Y = a_{02} + a_2 \cdot x_2 . \quad (16)$$

Мостеллер и Тьюки определяют новую переменную-отклик *Y**

$$Y^* = Y - a_{02} - a_2 \cdot x_2 , \quad (17)$$

которая получается из переменной *Y* исключением ее линейной связи с предиктором x_2. Далее из соотношения типа (7) для предиктора $x_1 = c_{01} + c_{12}x_2$ строится новая переменная x_1^*

$$x_1^* = x_1 - c_{01} - c_{12} \cdot x_2 , \quad (18)$$

которая представляет собой предиктор x_1, из которого также исключена его линейная зависимость от x_2; для построения *Y** (17) и x_1^* (18) использован метод наименьших квадратов.

Если теперь построить регрессионную модель вида

$$Y^* = a_{01}^* + a_1^* \cdot x_1^* , \quad (19)$$

которая описывает связи между *Y** и x_1^*, тогда имеет место равенство [8, с. 105]

$$b_1 = a_1^* . \quad (20)$$

Таким образом, коэффициент b_1 модели множественной регрессии (5) равен коэффициенту a_1^* модели простой регрессии (19), в которой использованы отклик *Y* и предиктор x_1, из которых удалена их линейная связь с предиктором x_2.

Соотношения (17)–(20) показывают ясный математический смысл коэффициентов МЛР с коррелированными предикторами. Тот факт, что речь идет о модели МЛР именно с коррелированными предикторами, подтверждается ненулевыми значениями коэффициентов c_{12} и c_{21} в соотношениях (7) и (18).

Другой вариант линейной корректировки описан в работе [14]. В обозначениях, использованных выше, показано, что для модели

$$Y = a_{01}^* + a_1^* x_1^* \quad (21)$$

выполняется то же соотношение (20). Отличие [14] от [8] заключается в том, что в первом случае для получения (20) использована сама переменная-отклик *Y*, а во втором – скорректированный отклик *Y** (17).

Соотношение (20) для моделей типа (21), написанное для самого отклика *Y* (не для *Y**), обобщается на любое число предикторов [15]. Для определенности рассмотрим коэффициент b_1 у предиктора x_1 в модели (1). Используя метод наименьших квадратов, построим регрессионную модель для предиктора x_1 по всем остальным предикторам x_j (аналог соотношения (7) для многих предикторов):

$$x_1 = c_{01} + c_{12}x_2 + c_{13}x_3 + \ldots + c_{1k}x_k \, .$$

Определим новую переменную x_1^* как

$$x_1^* = x_1 - c_{01} - c_{12}x_2 - c_{13}x_3 - \ldots . - c_{1k}x_k \, , \qquad (22)$$

которая представляет собой предиктор x_1 в модели (1), из которого исключена его линейная зависимость от всех остальных предикторов x_2, x_3, ..., x_k**.** Тогда коэффициент b_1 модели множественной регрессии (1) равен коэффициенту a_1^* модели простой регрессии (21), т. е. получаем соотношение (20). Это соотношение выполняется для любого предиктора x_i (i=1, 2, …, k)

$$b_i = a_i^* \, . \qquad (23)$$

Таким образом, коэффициент b_i модели (1) определяет влияние на отклик Y вовсе не предиктора X_i (как это утверждает «эпидемиологический» подход), а предиктора X_i^*. Возникает естественный вопрос: каков предметный смысл предиктора X_i^*, построенного по типу (22) для некоторого набора конфаундеров? Чем в предметном смысле отличается один X_i^*, построенный для одного набора конфаундеров, от такового, построенного для другого набора конфаундеров? Ни в одной известной нам публикации нет обсуждения предметного смысла (содержания) предиктора X_i^*.

Соотношения (17)–(23) определяют четкий математический (но не предметный) смысл коэффициентов b_i модели (1), и проблема *содержательной* (предметной) трактовки регрессионных коэффициентов b_i переносится на проблему содержательной интерпретации новых переменных x_i^* типа (22). Возможности предметно-ориентированной трактовки новых переменных x_i^* связаны с конкретной ситуацией; некоторые примеры приведены ниже.

Отметим, что оба *утверждения,* (15) и (23), справедливы только для аддитивной модели МЛР при условии линейной связи между откликом *Y* и предикторами с одной стороны, а также при условии линейных связей между самими предикторами X_i. Если модель не аддитивная и/или связь между предикторами нелинейная (произвольная), тогда трактовка коэффициентов моделей МЛР (1) будет иной.

Еще раз о смысле коэффициентов аддитивной модели множественной регрессии.

Интересно обсуждение моделей МЛР типа (1) с коррелированными предикторами, приведенное в монографии Н. Бейли [16, с. 187]. Регрессионная модель (5) с двумя предикторами у автора [16] называется «Частная регрессия». В этой модели автор [16] трактует коэффициенты b_1 и b_2 как частные коэффициенты регрессии по аналогии с частными коэффициентами корреляции (с. 188). Поскольку частный коэффициент корреляции между *Y* и x_1 получается исключением из *Y* и из x_1 их линейной зависимости от x_2 [7], логика трактовки коэффициентов регрессии в [16] получается похожей на описанную нами выше у Мостеллера и Тьюки (соотношения (17)–(20)). Таким образом, Н. Бейли [16] *понимал*, что коэффициенты b_1 и b_2 в МЛР с коррелированными предикторами не имеют «прямого» смысла показателя статистической связи отклика *Y* отдельно с X_1 и с X_2. Английская версия монографии Н. Бейли [16] опубликована в 1959 году и является самой ранней из известных нам, где коэффициенты МЛР называются не просто «коэффициенты регрессии», но «частные коэффициенты регрессии». Тем не менее в настоящее время среди эпидемиологов преобладает точка зрения [5] с формулами (3) и (4).

Термин «частные коэффициенты регрессии» употребляется также в монографии M. Kutner [et al.] [3]. Для модели с двумя предикторами авторы [3, P. 216] пишут: «Параметры b_1 и b_2 называют иногда *частными коэффициентами регрессии*, потому что они отражают частный эффект одного предиктора, когда другие предикторы…постоянны» (в оригинале: «The parameters b_1 and b_2 are sometimes called *partial regression coefficients* because they reflect the partial effect of one predictor variable when the other predictor variable is included in the model and is held constant»). Особенность этой формулировки заключается в том, что авторы *объясняют, почему* они называют коэффициенты частными – потому что при изменении одного предиктора другие предикторы удерживаются на постоянном уровне.

Сравним еще раз это определение с «эпидемиологическим» утверждением из монографии [5]: «коэффициент b_i… представляет собой изменение зависимой переменной *Y*, соответствующее изменению x_i на единицу, скорректированное на все другие члены в модели» (в оригинале дословно: «If the independent variable is a continuous variable such as x_i = age, then the interpretation of b_i is straightforward and represents the incremental change in the dependent variable, *Y* = body weight, associated with a unit change in x_i = age, adjusted for all other terms in the model. This is very much like the slope term in simple linear regression, so if b_{Age} = 2.0 kg, our interpretation is that the estimated body weight goes up 2.0 kg for each one year increment in age, adjusted for all the other terms in the model» [5, p. 77]. В этом *эпидемиологическом* утверждении нет ни слова о других предикторах, которые не изменяются при изменении предиктора x_i = age, но именно в этом ключе понимают коэффициенты b_i многие пользователи множественной регрессии, когда приводят в статьях значения некоторого b_i просто с перечислением тех остальных предикторов в МЛР, на которые был скорректирован данный коэффициент b_i (см., например, [6]).

*О возможной содержательной трактовке преобразованных предикторов X**

Пусть имеется аддитивная модель МЛР (1), в которой нас интересует влияние на отклик *Y* некоторого предиктора X_1. Рассмотрим несколько вариантов предметной трактовки предиктора X_1^*.

1) Предиктор X_1 и возможные конфаундеры имеют одну содержательную природу. В качестве примера рассмотрим связь общей заболеваемости *Y* населения Санкт-Петербурга с концентрациями загрязняющих веществ в атмосферном воздухе [17]. В работе [18] построена аддитивная модель

$$Y = 634{,}1 + 1047{,}1 \cdot CO - 278{,}4 \cdot SO_2, \quad (24)$$

где *CO* и SO_2 – концентрации оксида углерода и диоксида серы (оба предиктора в (24) представляют концентрации токсикантов в атмосферном воздухе). Согласно (24), изменение *Y* при изменении *CO* на единицу равно 1047,1 ед. Такое изменение *Y* при изменении *CO* возможно только при условии, что значения SO_2 остаются неизменными. Вместе с тем между *CO* и SO_2 имеется сильная корреляционная связь (r =0,729), и при изменении *CO* должны изменяться значения SO_2.

Построим для предиктора *CO* преобразованный предиктор *CO**. Используем для этого линейную связь между *CO* и SO_2, которая описывается соотношением *CO*=0,272 + 0,316 SO_2. Тогда предиктор *CO**, «скорректированный» на SO_2, имеет вид

$$CO^* = CO - 0{,}272 - 0{,}316 \cdot SO_2. \quad (25)$$

Связь заболеваемости *Y* с концентрацией *CO** имеет ожидаемый вид

$$Y \sim 1047{,}1 \cdot CO^*.$$

Это означает, что изменение Y при ΔCO =1, равное в (24) ΔY =1047,1 ед., справедливо в случае коррелированных предикторов CO и SO_2 только для преобразованного предиктора CO^* (25).

Специалисты-экологи и эпидемиологи, которые занимаются анализом состояния атмосферы городов и ее влияния на здоровье, наверное, могли бы как-то интерпретировать соотношение (25), в котором использована концентрация оксида углерода, из которой удалена ее связь с концентрацией диоксида серы SO_2.

Для модели (24) интересно обратиться к соотношению (13), согласно которому $\Delta Y(CO)$ определяется не коэффициентом b_1=1047,1 этой модели, а коэффициентом a_1 однофакторной модели связи Y только с одним предиктором CO. Согласно расчетам a_1=578,6, т. е. коэффициент a_1 связи Y с CO почти в два раза меньше, чем коэффициент b_1.

Какова же на *самом деле* связь заболеваемости Y с концентрацией CO, т. е. насколько изменяется Y при изменении CO? Связь Y с CO описывается соотношением Y~578,6·CO, т. е. заболеваемость изменяется на 578,6 ед. при изменении CO на единицу. Что касается соотношения Y~1047,1·CO^*, оно выполняется не для концентрации CO, а для преобразованной концентрации CO^*. Еще раз отметим, что все сказанное справедливо только в случае *линейной* связи Y с CO и линейной связи CO с SO_2.

2) Предиктор X_1 считается скорректированным на «традиционные» для эпидемиологии конфаундеры, такие как возраст и пол.

$$Y=b_0 + b_1\cdot CO + b_2\cdot\text{Возраст} + b_3\cdot\text{Пол}. \qquad (26)$$

Такого рода корректировки для эпидемиологии являются традиционными (см., например, [6, 19]). Авторы [6, 19], следуя эпидемиологическому подходу [5], предполагают, видимо, что с помощью соотношения (26) влияние CO на заболеваемость жителей разного возраста и пола уравнивается (стандартизуется) по возрасту и полу. Конечно, это неверно, но какая-то логика в таком предположении есть.

3) Модель МЛР включает предиктор X_1 и конфаундеры различной природы, например,

$$Y=b_0 + b_1\cdot CO + b_2\, SO_2 + b_3\cdot\text{Возраст} + b_4\cdot\text{Пол}.$$

Теперь концентрация CO^* получается из CO путем удаления из нее линейной связи с SO_2, а также с Возрастом и Полом. Получающийся «конгломерат» CO^* едва ли поддается какой-то предметной интерпретации.

4. Регрессионные модели, нелинейные по предикторам

Рассмотренные выше модели, линейные по предикторам (аддитивные модели), не учитывают многих особенностей фактического эпидемиологического материала, например, нелинейные связи некоторых предикторов X_i с откликом Y, а также корреляции между предикторами и возможное наличие неаддитивных эффектов (отклонение совместного влияния нескольких предикторов на Y от простого суммирования). С этих позиций рассмотрим несколько простых примеров МЛР, нелинейных по предикторам.

4.1. Модель, квадратичная по предиктору X

Пусть имеется модель ПЛР типа (2) с одним предиктором $Y(x)=a_0+a_1x$, в которой коэффициент a_1 показывает, что при изменении X на единицу, изменение Y имеет вид, аналогичный (4): $\Delta Y(x,\Delta x=1)=Y(x+1)-Y(x)=a_1$. Это изменение

$\Delta Y(x, \Delta x = 1)$ определяется одним коэффициентом a_1. Предположим, что связь Y с X является нелинейной и, более конкретно, описывается полиномом второй степени по X вида

$$Y(x) = a_0 + a_1 x + a_2 x^2 . \quad (27)$$

Здесь при изменении X на единицу будем иметь следующую формулу для расчета изменения Y:

$$\Delta Y(x, \Delta x = 1) = Y(x+1) - Y(x) = a_1 + a_2(2x+1) . \quad (28)$$

Теперь влияние X на Y нельзя описать одним или двумя коэффициентами a_1 и a_2, поскольку изменение Y зависит не только от этих коэффициентов, но и от того значения предиктора x, при котором рассчитывается $\Delta Y(x, \Delta x = 1)$. При этом в модели (27) имеется всего *один* предиктор x в различных степенях. Если оставаться на позициях аддитивных моделей, которые при оценке $\Delta Y(x, \Delta x = 1)$ опираются только на регрессионные коэффициенты, едва ли кому-то придет в голову трактовать коэффициент a_1 как показатель *влияния* предиктора x на Y, *скорректированного* на предиктор x^2.

Этот простой пример показывает, что в нелинейных по предикторам моделях оценка влияния предиктора X_i на Y не может опираться только на коэффициенты модели, как это предлагается в моделях, линейных по предикторам; оценка должна основываться на всем соотношении (27), описывающем связь Y с X.

4.2. Модель с перекрестным членом

Рассмотрим другой пример модели, нелинейной по предикторам – модель с перекрестным членом. Пусть имеем аддитивную модель типа (1), содержащую два предиктора x_1 и x_2. Добавим в эту аддитивную модель перекрестный член, в котором значения предикторов x_1 и x_2 перемножаются [1–4]:

$$Y(x_1, x_2) = b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2 . \quad (29)$$

Оценка отклика $Y(x_1, x_2)$ так же, как аналогичная оценка $Y(x)$ в (27), должна основываться на всей функции (27), а не только на коэффициентах модели. В отличие от квадратичной модели (27) для одного предиктора x, здесь при расчете изменения Y при изменении x_1 или x_2 еще необходимо учитывать корреляции между предикторами x_1 и x_2. Корреляции приводят к тому, что при изменении, например, предиктора x_1, необходимо учитывать также изменение предиктора x_2, и модель (29), в отличии от (1), позволяет это сделать. В модели (29) изменение Y при изменении x_1 имеет вид:

$$\Delta Y(x_1, \Delta x_1) = Y(x_1 + \Delta x_1, x_2 + \Delta x_2) - Y(x_1, x_2) = b_1 \Delta x_1 + b_2 \Delta x_2 + b_{12}(x_1 \Delta x_2 + x_2 \Delta x_1 + \Delta x_1 \Delta x_2) , \quad (30)$$

где Δx_2 – это изменение предиктора x_2 за счет корреляции между x_1 и x_2. Таким образом, изменение отклика Y при изменении предиктора x_1 зависит, согласно (30), не только от величины Δx_1, но и от Δx_2, и от значений обоих предикторов x_1 и x_2, при которых рассчитывается $\Delta Y(x_1, \Delta x_1)$.

5. Определение однофакторных функций $Y=Y(x_i)$

Из предыдущих разделов следует вывод, что в общем случае модель МЛР должна быть нелинейной по предикторам, т. е. включать значимые члены более

высокого порядка по x_i, чем линейные, а также перекрестные члены. Если по минимуму ограничиться квадратичными членами по x_i и парными перекрестными членами ($x_i \cdot x_j$), получаем модель МЛР вида

$$Y(x_1, x_2, ..., x_k) = b_0 + \sum_{i=1}^{k} b_i x_i + \sum_{i=1}^{k} b_{ii} x_i^2 + \sum_{i=1}^{k} \sum_{j>i}^{k} b_{ij} x_i x_j + \varepsilon, \quad (31)$$

где ε – ошибки интерполяции экспериментальных значений отклика *Y* функцией (31). Методом наименьших квадратов определим коэффициенты $\hat{b}_i$ модели МЛР (31), которые обычно изображают символами с крыжом. Используем также какие-либо шаговые процедуры, которые помогут отсеять статистически незначимые слагаемые в (31). Полученное таким образом выражение для расчетной (*calculated*) функции отклика *Y* будем обозначать Y_{calc} (x_1, x_2, …, x_k)

$$Y_{calc}(x_1, x_2, ..., x_k) = \hat{b}_0 + \sum_{i} \hat{b}_i x_i + \sum_{i} \hat{b}_{ii} x_i^2 + \sum_{i} \sum_{j} \hat{b}_{ij} x_i x_j, \quad (32)$$

где в выражении (32) оставлены только те слагаемые, которые оказались статистически значимы.

Как отмечалось, влияние каждого предиктора x_i на отклик *Y* определяется не коэффициентами $\hat{b}_i$ модели (32), а всем соотношением Y_{calc} (x_1, x_2, …, x_k). Допустим, нас интересует однофакторная функция $Y(x_1)$ некоторого выбранного предиктора x_1. Для нахождения $Y(x_1)$ при наличии *n*-объектов исследования производим расчет Y_{calc} (x_1, x_2, …, x_k) по формуле (32) с использованием ***наблюдаемых*** значений ***всех*** предикторов x_i. В результате получаем *n*-значений *Y*, каждому из которых соответствует набор значений предикторов x_1, x_2, …, x_k. Использование именно наблюдаемых значений x_i позволяет учесть все корреляции между предикторами, которые оказывают влияние на вид Y_{calc} (x_1, x_2, …, x_k). Тогда однофакторная функция предиктора x_1 представляется *n* расчетными значениями *Y* (*n* числами), каждому из которых поставлено в соответствие значение одного предиктора x_1

$$Y(x_1) = Y_{calc}(x_1). \quad (33)$$

Здесь Y_{calc} (x_1) – это функция Y_{calc} (x_1, x_2, …, x_k), в которой нас интересует ее зависимость только от x_1.

Итог: что такое однофакторная функция $Y=Y(x_i)$ конкретного предиктора x_i в неаддитивной модели множественной линейной регрессии $Y(x_1, x_2, …, x_k)$ с учетом всех остальных предикторов x_j при возможном наличии корреляций между ними? Ответ выглядит следующим образом:

– в числовом виде $Y=Y(x_i)$ – это два набора *n*-чисел: набор Y_{calc} (x_1, x_2, …, x_k), получаемых путем расчетов по формуле (32) при использовании всех предикторов x_1, x_2, …, x_k, и набор значений предиктора x_i независимо от остальных предикторов в модели (32).

– в графическом виде $Y=Y(x_i)$ – это диаграмма рассеяния в координатах Y_{calc} (x_1, x_2, …, x_k) и x_i.

– в аналитическом (формульном) виде $Y=Y(x_i)$ – это результат интерполяции точек на диаграмме рассеяния некоторой функцией Y_{int}, по возможности «гладкой» (int – это *interpolation*). Поскольку в реальных задачах разброс точек на диаграммах рассеяния обычно оказывается велик, для получения именно «гладкой» функции Y_{int} предлагается использовать методы скользящего среднего [8, 19].

Еще один вариант определения однофакторной функции $Y=Y(x_i)$ получается следующим образом. Допустим, нам удалось построить модель МЛР (32) очень

высокого качества, когда коэффициент детерминации модели R^2 приближается к единице. В этом случае расчетные значения отклика Y_{calc} (x_1, x_2, …, x_k) совпадают с наблюдаемыми (*observed*) значениями Y_{obs}. Тогда в такой модели для выбранного предиктора x_1 функция $Y(x_1) = Y_{calc}(x_1) = Y_{obs}(x_1)$, или просто

$$Y(x_1) = Y_{obs}(x_1), \tag{34}$$

где n расчетных, как в формуле (33), значений Y_{calc} заменены на n наблюдаемых значений отклика Y_{obs}.

В случае соотношения (34) однофакторная функция $Y=Y(x_i)$ предиктора x_i определяется совершенно аналогично случаю соотношения (33): в числовом виде $Y=Y(x_i)$ – это два набора n-чисел Y_{obs} и x_i в базе данных, а в формульном виде – это интерполяционная функция Y_{int}, полученная интерполяцией по значениям Y_{obs} и x_i. Второй вариант (34) определения однофакторной функции $Y=Y(x_i)$, основанный на наблюдаемых значениях отклика Y_{obs}, выглядит более универсальным, чем первый (33), поскольку не требует построения модели множественной регрессии, а базируется лишь на наблюдаемых значениях Y_{obs} и значениях конкретного x_i.

Соотношение (33) переходит в (34) при условии, что коэффициент детерминации $R^2 \to 1$. Что будет, если R^2 конкретной модели МЛР далек от единицы? Это означает, что в модель не были включены какие-то предикторы, значимые для предсказания отклика Y. Это могут быть предикторы, известные или неизвестные нам; в любом случае, их гипотетическое включение должно привести к R^2, близкому к единице. Поскольку соотношение $Y(x_1) = Y_{obs}(x_1)$ выполняется вне зависимости от состава предикторов в МЛР, мы опять приходим к соотношению (34).

Подчеркнем, что в обоих вариантах, (33) и (34), определения однофакторных функций $Y=Y(x_i)$ нет никакой надобности в «скорректированных на конфаундеры» коэффициентах b_i; термин «корректировка» вообще не фигурирует в формулах (33) и (34). На фоне предложений, высказанных эпидемиологами в монографии [5], этот вывод кажется достаточно «радикальным», но здесь мы не одиноки. Например, Э. Сигел пишет: «*Какой была бы эта взаимосвязь* (взаимосвязь Y с одной переменной X_i. – *Вставка наша*)*, если бы остальные переменные... не фиксировались на постоянном уровне? На этот вопрос можно ответить, проанализировав... коэффициент регрессии, прогнозирующий Y на основании только одной этой X-переменной*" [2, с. 621]. Поскольку у Э. Сигела речь идет о коэффициенте регрессии для связи Y с одной переменной, значит, в упомянутой цитате – это коэффициент модели *простой* регрессии a_i, а вовсе не коэффициент b_i в модели *множественной* регрессии (1). Далее, поскольку речь в цитате [2] идет о регрессионных коэффициентах, очевидно, предполагается, что связь Y с предиктором x_i является линейной и описывается аддитивной моделью (1). В общем случае неаддитивных моделей взаимосвязь Y с предикторами с учетом корреляций между ними описывается не коэффициентами МЛР, а соотношениями типа (32). При этом однофакторные функции $Y=Y(x_i)$ для связи Y с одним предиктором x_i, когда остальные предикторы не фиксируются на постоянном уровне, описываются соотношениями (33) и (34).

6. Соотношение методов стандартизации и регрессии

Стандартизация путем стратификации конфаундеров позволяет уравнять объекты исследования по тем конфаундерам, которые являются значимыми для данной задачи. Представим гипотетическую ситуацию, когда отклик Y связывается статистической связью с двумя предикторами X_1 и X_2. Оба предиктора имеют статистически значимые связи с Y; имеется также значимая корреляционная связь между X_1 и X_2. Можно написать уравнение регрессии, связывающее Y с X_1:

$$Y = a_0 + a_1 x_1 . \quad (35)$$

Если в уравнение регрессии (35) добавить второй предиктор, получим

$$Y = b_0 + b_1 x_1 + b_2 x_2 . \quad (36)$$

Поскольку предикторы X_1 и X_2 по условию задачи коррелируют, тогда $b_1 \neq a_1$. Вопрос о причинах корреляций между X_1 и X_2 оставим за скобками. Нас интересует другие вопросы: 1) какой из коэффициентов – a_1 или b_1 – «правильно» отражает влияние предиктора X_1 на отклик Y? 2) как в модели (36) учесть совместное влияние на Y предикторов X_1 и X_2?

Согласно доказательствам, представленным выше, регрессионная модель (36) не может ответить сразу *на оба* вопроса. Ответ на первый вопрос: если связь Y с X_1 действительно линейная, тогда силу связи Y с X_1 при наличии различных значений X_2 описывает ***в среднем*** коэффициент a_1 модели (35). Ответ на второй вопрос: если необходимо оценить в среднем силу связи Y с X_2, тогда необходимо построить уравнение простой регрессии типа (35), которая связывает Y с одним предиктором X_2. Если все же требуется описать связь Y одновременно с X_1 и с X_2, модель (36) при наличии корреляций между X_1 и X_2 оказывается бесполезной: коэффициенты b_1 и b_2 модели (36) не описывают эту связь.

Что можно сделать для описания связи Y с X_1 при различных X_2, если использовать только линейные по предикторам модели? Здесь возможны следующие варианты действий, основанные на стратификации предикторов.

1. Значения предиктора X_2 разделяют в простейшем случае на два диапазона (две страты), из которых одна страта содержит низкие значения X_2, а вторая – высокие X_2. Разделения X_2 на страты могут быть проведены по различным критериям, например, по медиане. Для такого разделенного массива наблюдений можно построить две регрессионные модели типа (35), в одной из которых использованы значения Y и X_1, соответствующие низким значениям X_2, а в другой – значения тех же Y и X_1 при высоких X_2. Сила влияния X_1 на Y будет в обоих случаях определяться коэффициентами a_1 этих двух регрессионных моделей. Различия этого подхода от описанного выше для всего (неразделенного) массива наблюдений заключаются

в том, что вместо одного коэффициента a_1 имеем два коэффициента, каждый из которых описывает связь Y с X_1 при различных значениях X_2. Если эти два коэффициента a_1 будут различными, значит, влияние X_1 на Y зависит от X_2. Таким образом, влияние X_2 включается в анализ связи Y с X_1. При необходимости (возможности) значения X_2 разделяются на произвольное (большее двух) число страт.

2. На два (в простейшем случае) диапазона разделяют значения обоих предикторов X_1 и X_2: условно низкие (1) и высокие (2) значения предикторов. В результате получается четыре набора данных (см. табл.).

Таблица. Наборы данных

X_2 – уровни	X_1 – уровни	
	1	2
1	Y (11), X_1 (11), X_2 (11), n_{11}	Y (12), X_1 (12), X_2 (12), n_{12}
2	Y (21), X_1 (21), X_2 (21), n_{21}	Y (22), X_1 (22), X_2 (22), n_{22}

Набор, состоящий из n_{11} значений отклика и предикторов, содержит данные с низкими значениями обоих предикторов X_1 (11) и X_2 (11), а также соответствующие

им значения отклика *Y* (11). Аналогично понимается содержание остальных ячеек табл. 1.

После такой стратификации возможны два варианта решений.

1) Если числа наблюдений n_{ij} достаточно велики, можно построить уравнения простой регрессии для данных каждой ячейки табл. 1. Например, уравнение типа $Y(11) = a_0 + a_1(11)x_1(11)$ будет описывать связь *Y* с X_1 при низких значениях предикторов X_1 и X_2. Сила связи *Y* с X_1 будет описываться (при условии линейности связи *Y* с X_1 для набора n_{11}) коэффициентом $a_1(11)$. Возможные различия коэффициентов a_1(i,j) покажут различия влияния X_1 на *Y* при различных X_1 и X_2.

2) Если оказывается, что при разделении предикторов X_1 и X_2 на две (или более) страты числа наблюдений n_{ij} в некоторых ячейках табл. 1 стали так малы, что построить регрессионную модель невозможно, то для оценки (хотя бы грубой) связи *Y* с X_1 при различных X_2 можно использовать средние значения *Y*(ij) в соответствующих ячейках табл. 1. Для этого, правда, все равно нужны значения n_{ij}, превышающие n_{ij}=1 (для расчета среднего значения нужны как минимум два наблюдения). Тогда средние значения *Y*(ij), представленные на графике типа *Interaction Plots* в дисперсионном анализе, покажут тенденции изменения *Y* при изменении предикторов в выбранных диапазонах *i,j*.

Эпидемиологическое предложение использовать регрессию вместо стандартизации вызвано именно тем, что при большом числе конфаундеров число наблюдений n_{ij} в некоторых стратах может стать нулевыми. При наличии нулевых n_{ij} провести стандартизацию оказывается невозможно; при этом неявно считается, что регрессия в виде (36) может быть построена при любых n_{ij} (да, может). Однако для применения регрессии вместо стандартизации необходимо иметь доказательства их эквивалентности в деле по устранению конфаундинга. Никаких доказательств этого нет и быть не может, поскольку два метода (стандартизация и регрессия) совсем разные.

7. Выводы

В эпидемиологическом подходе, принятом Всемирной организацией здравоохранения [5], предполагается, что регрессионное моделирование используют для корректировки на конфаундеры как эквивалент процедуре стандартизации путем стратификации конфаундеров. Аналогичной точки зрения придерживаются известные российские эпидемиологи [20]. Никаких доказательств этой эквивалентности в [5, 20] не представлено. В настоящей работе показано, что статистическое моделирование методами линейной регрессии не эквивалентно процедуре стандартизации.

Процедура Adjustment for confounders, приводящая для аддитивных моделей МЛР к соотношениям (3) и (4), не имеет математического обоснования и не может использоваться для оценки влияния каждого предиктора x_i среди множества других предикторов x_j в модели множественной линейной регрессии $Y(x_1, x_2, x_3, \ldots, x_k)$. Вместо соотношений (3) и (4) для оценки однофакторных эффектов $Y(x_i)$ должно использоваться соотношение (33) или (34); могут быть также полезны подходы с использованием стратификации предикторов, как это описано в разделе 6.

Даже если допустить, что в *эпидемиологическом* утверждении «b_i – это эффект предиктора x_i, скорректированный на конфаундеры» есть *доля истины*, тогда все равно должны быть учтены следующие положения:

1) коэффициент b_i у предиктора x_i в аддитивной модели (1) зависит от набора *сопутствующих* предикторов x_j (конфаундеров), которые включены в модель помимо предиктора x_i. Мостеллер и Тьюки в монографии [8] неоднократно подчеркивают, что при различных наборах конфаундеров x_j коэффициент b_i имеет *различные* значения.

В этом случае при каком наборе конфаундеров x_j значение коэффициента b_i следует

считать «правильным», «истинным»? Такую ситуацию мы видим во многих публикациях. Например, в [6] построены три модели (одна без конфаундеров и две с разными наборами конфаундеров) и получены различные коэффициенты b_i. Авторы [6] не сообщают, какой из трех коэффициентов следует использовать для оценки реальных эффектов.

2) для использования формы $Y(x_i) \sim b_i x_i$ необходимо, чтобы связь Y с x_i не просто *предполагалась* линейной, как в модели типа (1), но и *была в реальности* линейной. Авторы, использующие эпидемиологический подход, часто не приводят в статьях диаграммы рассеяния (Y – x_i), не проверяют линейность связи $Y(x_i)$ с x_i другими способами и/или не сообщают о проводимой проверке [6, 21].

3) апелляция к «корректировке на конфаундеры» в стиле $\Delta Y(\Delta x_i = 1) = b_i$ в принципе невозможна для неаддитивных моделей, которые являются более общими, чем аддитивные модели. В любой неаддитивной модели имеется *как минимум* два слагаемых, которые содержат предиктор x_i и, как следствие, два регрессионных коэффициента b_i. Апеллировать к одному из коэффициентов, скорректированному на конфаундеры, становится принципиально невозможно, поэтому для оценки величины однофакторных эффектов $\Delta Y(\Delta x_i = 1)$ необходимо использовать подход (33) или (34).

8. Благодарности

Исследование выполнено за счет субсидий Минобрнауки РФ на выполнение научной темы FUMN-2024-0002.

9. Список литературы

1. *Draper, N. R.* Applied Regression Analysis / N. R. Draper, H. Smith. – 3rd ed. – Hoboken : Wiley Interscience Publication, 1998. – 736 p.
2. *Сигел, Э.* Практическая бизнес-статистика / Э. Сигел. – Москва : И. Д. Вильямс, 2017. – 1056 с. – ISBN 978-5-8459-1367-8.
3. Applied linear statistical models / M. H. Kutner, C. J. Nachtsheim, J. Neter, W. Li. – 5th ed. – New York : McGraw-Hill/Irwin, 2005. – 1396 p.
4. *Greenland, S.* Interactions in Epidemiology: Relevance, Identification, and Estimation / S. Greenland // Epidemiology. – 2009. – Vol. 20, № 1. – P. 14–17. – DOI 10.1097/EDE.0b013e318193e7b5.
5. *Bonita, R.* Basic Epidemiology / R. Bonita, R. Beaglehole, T. Kjellström. – 2nd ed. – Geneva : World Health Organization, 2006. – 212 p.
6. *Yang, B. R.* Electronic cigarette use and heavy metal exposure: Evidence from the Korea National Health and Nutrition Examination Survey / B. R. Yang, S.-M. Lee // Annals of Epidemiology. – 2025. – Vol. 103. – P. 1–8. – DOI 10.1016/j.annepidem.2025.01.011.
7. *Afifi, A. A.* Computer-aided multivariate analysis / A. A. Afifi, S. May, V. Clark. – 4th ed. – London : Chapman&Hall/CRC, 2003. – 512 p. – ISBN 978-1584883081.
8. *Мостеллер, Ф.* Анализ данных и регрессия / Ф. Мостеллер, Дж. Тьюки. – М. : Финансы и статистика, 1982. – 378 с.
9. *Ehrenberg, A. S. C.* Data reduction / A. S. C. Ehrenberg. – Chichester : John Wiley&Sons Inc, 1975.
10. *Abbott, R. D.* Interpreting multiple logistic regression coefficients in prospective observational studies / R. D. Abbott, R. J. Carroll // Am. J. Epidemiol. – 1984. – Vol. 119, № 5. – P. 830–836.
11. *Маслакова, Т. А*. Интерпретация прогностических регрессионных моделей в области медико-экологического мониторинга / Т. А. Маслакова, А. Н. Вараксин, В. Н. Чуканов // Экологические системы и приборы. – 2008. – № 2. – С. 6–9.
12. *Панов, В. Г.* Соотношение между моделями простой и множественной регрессии в задачах медико-экологического мониторинга / В. Г. Панов, А. Н. Вараксин, Т. А. Маслакова // Экологические системы и приборы. – 2008. – № 7. – С. 15–17.

13. *Panov, V. G.* Relation between the coefficients of simple and multiple regression models / V. G. Panov, A. N. Varaksin // Siberian Mathematical Journal. – 2010. – Vol. 51, No. 1. – P. 162–167.
14. *Вараксин, А. Н.* О коэффициентах уравнения множественной регрессии с коррелированными предикторами в задачах медико-экологического мониторинга / А. Н. Вараксин, В. Г. Панов // Экологические системы и приборы. – 2011. – № 4. – С. 29–33.
15. *Panov, V. G.* Relationships between linear statistical models / V. G. Panov, A. N. Varaksin // British Journal of Mathematics & Computer Science. – 2015. – Vol. 11, № 6. – P. 1–18. – DOI 10.9734/BJMCS/2015/20493.
16. *Бейли, Н.* Статистические методы в биологии / Н. Бейли ; перевод с английского С. Я. Зайцева. – М. : Иностранная литература, 1962. – 260 с.
17. *Щербо, А. П.* Окружающая среда и здоровье: подходы к оценке риска / А. П. Щербо. – Санкт-Петербург : СПбМАПО, 2002. – 376 с.
18. *Вараксин, А. Н.* Регрессионные модели в задачах оценки риска здоровью: алгоритм учета корреляций между предикторами / А. Н. Вараксин // Экологические системы и приборы. – 2025. – № 1. – С. 3–14. – DOI 10.25791/esip.1.2025.1495
19. Применение методов скользящего среднего для построения регрессионных моделей в медико-экологических исследованиях / А. Н. Вараксин, Ю. В. Шалаумова, Т. А. Маслакова [и др.] // Экологические системы и приборы. – 2020. – № 6. – С. 12–21. – DOI 10.25791/esip.06.2020.1159. – EDN XTBFAV.
20. *Tinajero, M. G.* Evaluation of ethnic differences in cardiometabolic risk in children / M. G. Tinajero, C. Keown-Stoneman, V. S. Malik // Annals of Epidemiology. – 2023. – Vol. 86. – P. 121–126. – DOI 10.1016/j.annepidem.2023.05.014.

Сведения об авторе:

Вараксин Анатолий Николаевич, д. ф.-м. н., профессор, г. н. с. лаборатории биостатистики Института промышленной экологии, г. Екатеринбург, Россия. Эл. почта: varaksin@ecko.uran.ru.

ON THE CONCEPT OF "ADJUSTMENT FOR CONFOUNDERS" IN LINEAR REGRESSION MODELS

A. N. Varaksin

Institute of Industrial Ecology of UB RAS, Ekaterinburg, Russian Federation

This methodological article addresses issues arising in interpreting the results of multiple linear regression (MLR) with correlated predictors. Correlation between predictors X leads to the multiple regression coefficients, which describe the impact of a certain set of predictors on the response Y, varying depending on the set of predictors included in the MLR model. In particular, they differ from the coefficients of simple regression models, each of which describes the impact of one predictor on the response. Simple regression coefficients are commonly referred to as "crude," while MLR coefficients are referred to as "adjusted" for the remaining predictors included in the MLR. This approach is called "adjustment for confounders." This paper demonstrates that the effect of any predictor X_i on the response Y is described not by the regression coefficient b_i for this predictor, but by the entire regression function $Y=Y(x_1, x_2, ... , x_k)$, which depends on all predictors $x_1, x_2, ... , x_k$. Analysis of this expression suggests that the influence of the predictor X_i on the response Y is ultimately described by a one-factor function $Y=Y(x_i)$, constructed from the calculated $Y_{calc} (x_1, x_2, ... , x_k)$ or observed Y_{obs} values of the response Y and the values of the given predictor x_i. It is demonstrated that the "adjustment for confounders" procedure lacks mathematical justification and cannot be used in the analysis of multiple linear regression models.

Key words: multiple linear regression; additive and non-additive models; correlated predictors; adjustment for confounders; one-factor response functions.

References

1. *Draper, N. R.* Applied Regression Analysis / N. R. Draper, H. Smith. – 3rd ed. – Hoboken : Wiley Interscience Publication, 1998. – 736 p.
2. *Sigel, E*. Practical business statistics / E. Sigel. – Moscow: I. D. Williams LLC, 2017. – 1056 p. – ISBN 978-5-8459-1367-8. (in Russian).
3. Applied linear statistical models / M. H. Kutner, C. J. Nachtsheim, J. Neter, W. Li. – 5th ed. – New York : McGraw-Hill/Irwin, 2005. – 1396 p.
4. *Greenland, S.* Interactions in Epidemiology: Relevance, Identification, and Estimation // Epidemiology. – 2009. – Vol. 20, № 1. – P. 14–17. – DOI 10.1097/EDE.0b013e318193e7b5.
5. *Bonita, R.* Basic Epidemiology / R. Bonita, R. Beaglehole, T. Kjellström. – 2nd ed. – Geneva : World Health Organization, 2006. – 212 p.
6. *Yang, B. R.* Electronic cigarette use and heavy metal exposure: Evidence from the Korea National Health and Nutrition Examination Survey / B. R. Yang, S.-M. Lee // Annals of Epidemiology. – 2025. – Vol. 103. – P. 1–8. – DOI 10.1016/j.annepidem.2025.01.011.
7. *Afifi, A. A.* Computer-aided multivariate analysis / A. A. Afifi, S. May, V. Clark. – 4th ed. – London : Chapman&Hall/CRC, 2003. – 512 p. – ISBN 978-1584883081.
8. *Mosteller, F.* Data Analysis and Regression / F. Mosteller, J. Tukey – Moscow : Finance and statistics, 1982. – 378 p. (in Russian).
9. *Ehrenberg, A. S. C.* Data reduction / A. S. C. Ehrenberg. – Chichester : John Wiley&Sons Inc, 1975.
10. *Abbott, R. D.* Interpreting multiple logistic regression coefficients in prospective observational studies / R. D. Abbott, R. J. Carroll // Am. J. Epidemiol. – 1984. – Vol. 119, № 5. – P. 830–836.
11. *Maslakova, T. A.* Interpretation of predictive regression models in the field of medical and environmental monitoring / T. A. Maslakova, A. N. Varaksin, V. N. Chukanov // Ecological systems and devices. – 2008. – № 2. – P. 6–9. (in Russian).
12. *Panov, V. G.* Relationship between simple and multiple regression models in medical and environmental monitoring tasks / V. G. Panov, A. N. Varaksin, T. A. Maslakova // Ecological systems and devices. – 2008. – № 7. – P. 15–17. (in Russian).

13. *Panov, V. G.* Relation between the coefficients of simple and multiple regression models / V. G. Panov, A. N. Varaksin // Siberian Mathematical Journal. – 2010. – Vol. 51, No. 1. – P. 162–167.
14. *Varaksin, A. N.* On the coefficients of the multiple regression equation with correlated predictors in problems of medical and ecological monitoring / A. N. Varaksin, V. G. Panov // Ecological systems and devices. – 2011. – № 4. – P. 29–33. (in Russian).
15. *Panov, V. G.* Relationships between linear statistical models / V. G. Panov, A. N. Varaksin // British Journal of Mathematics & Computer Science. – 2015. – Vol. 11, № 6. – P. 1–18. – DOI 10.9734/BJMCS/2015/20493.
16. *Bailey, N.* Statistical Methods in Biology / N. Bailey – Moscow : Foreign Literature Publishing House, 1962 – 260 p. (in Russian).
17. *Shcherbo, A. P.* Environment and health: approaches to risk assessment / A. P. Shcherbo. – St. Petersburg: St. Petersburg Medical Academy of Postgraduate Education, 2002. – 376 p. (in Russian).
18. *Varaksin, A. N.* Regression models in health risk assessment problems: an algorithm for taking into account correlations between predictors / A. N. Varaksin // Ecological systems and devices – 2025. – № 1. – P. 3–14. – DOI 10.25791/esip.1.2025.1495 (in Russian).
19. Application of moving average methods for constructing regression models in medical and ecological research / A. N. Varaksin, Yu. V. Shalaumova, T. A. Maslakova [et al.] // Ecological systems and devices. – 2020. – № 6. – P. 12–21. – DOI 10.25791/esip.06.2020.1159. – EDN XTBFAV. (in Russian).
20. *Tinajero, M. G.* Evaluation of ethnic differences in cardiometabolic risk in children / M. G. Tinajero, C. Keown-Stoneman, V. S. Malik // Annals of Epidemiology. – 2023. – Vol. 86. – P. 121–126. – DOI 10.1016/j.annepidem.2023.05.014.