

Статьи: эконометрическая теория

Некоторые эквивалентности в линейном оценивании*

Дмитрий Данилов[†]

Эйндховенский технологический университет, Эйндховен, Нидерланды

Ян Магнус[‡]

Тилбургский университет, Тилбург, Нидерланды

В условиях нормальности задача байесовского оценивания, задача наилучшего линейного несмещенного оценивания и задача наименьших квадратов при ограничениях эквивалентны. В результате нет необходимости рассчитывать псевдообратные матрицы и выполнять другие сложные операции, невозможные для больших разреженных систем. Вместо этого, преобразовав входные показатели, можно переписать систему в виде эквивалентной, которую уже можно разрешить обычным методом наименьших квадратов.

Линейное байесовское оценивание, наилучшее линейное несмещенное оценивание, наименьшие квадраты, разреженные задачи, крупномасштабная оптимизация
Классификация JEL: C11, C61, C63

1 Введение

Пусть дан вектор y размерности $n \times 1$ и матрица X размерности $n \times k$ с линейно независимыми столбцами. Подразумевается, что вектор y и матрица X известны (и неслучайны). Задача состоит в том, чтобы определить вектор β размерности $k \times 1$, удовлетворяющий уравнению

$$y = X\beta.$$

Пусть $M := I_n - X(X'X)^{-1}X'$ – привычная идемпотентная матрица. Если $My = 0$, то уравнение $y = X\beta$ имеет единственное решение

$$\hat{\beta} := (X'X)^{-1}X'y.$$

Если $My \neq 0$, то уравнение не имеет решений. В этом случае можно найти вектор $\hat{\beta}$, который в некотором смысле минимизирует вектор «ошибки» $e = y - X\beta$. Удобная скалярная мера ошибки –

$$e'e = (y - X\beta)'(y - X\beta),$$

и известно, что $\hat{\beta}$ минимизирует $e'e$ по всем k -мерным β -векторам. Вектор $\hat{\beta}$ называется *решением наименьших квадратов*, а $X\hat{\beta}$ – *приближением наименьших квадратов* для y . Таким образом, $\hat{\beta}$ является «наилучшим» выбором для β независимо от того, совместно уравнение $y = X\beta$ или нет. Если $y = X\beta$ совместно, то $\hat{\beta}$ является решением; если $y = X\beta$ несовместно, то $\hat{\beta}$ – решение наименьших квадратов.

*Перевод С. Анатольева. Цитировать как: Данилов, Дмитрий и Ян Магнус (2007). «Некоторые эквивалентности в линейном оценивании», Квантиль, №3, стр. 83–90. Citation: Danilov, Dmitry and Jan R. Magnus (2007). “Some equivalences in linear estimation,” Quantile, No.3, pp. 83–90.

[†]Адрес: Eindhoven University of Technology, P.O. Box 513, 5600 MB, Eindhoven, The Netherlands. Электронная почта: danilov@eurandom.tue.nl

[‡]Адрес: CentER and Department of Econometrics & OR, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Электронная почта: magnus@wt.nl

В отличие от детерминистической задачи наименьших квадратов, стандартная задача линейной регрессии формулируется в стохастическом контексте, где рассматривается

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n),$$

откуда выводятся оценка Гаусса–Маркова и ее дисперсионная матрица как

$$\hat{\beta} = (X'X)^{-1}X'y, \quad \mathbb{V}(\hat{\beta}) = \sigma^2(X'X)^{-1},$$

см., например, Магнус, Катышев & Пересецкий (2005). Тот факт, что решение наименьших квадратов и оценка Гаусса–Маркова идентичны, ни в коей мере не является очевидным. Он был тщательно изучен и обозначен в Rao (1971, 1973). Эта эквивалентность привела к неудачному использованию термина «(обычной) оценки наименьших квадратов», означающего оценку Гаусса–Маркова. Метод наименьших квадратов, однако, – это чисто детерминистический метод, завязанный на приближении, а не на оценивании.

В отличие от классического (частотного) подхода, байесовский не предполагает «истинных» β -параметров. Вместо этого предполагается вероятностное распределение параметров, так называемое априорное распределение. Данные помогают модифицировать априорное знание «правды» в более законченное знание – апостериорное распределение. Формула, порождающая этот переход, – это формула Байеса

$$p(\beta|y) = \frac{\pi(\beta)p(y|\beta)}{p(y)},$$

где $\pi(\beta)$ обозначает априорное распределение, $p(y|\beta)$ – привычная функция правдоподобия, $p(\beta|y)$ – апостериорное распределение, а $p(y)$ – коэффициент пропорциональности, возникающий из-за того, что $p(\beta|y)$ должно интегрироваться в единицу. Например, если мы (на время) предположим, что Σ известно, и что все распределения нормальны, то байесовский статистик использует данные

$$y|\beta \sim N(X\beta, \Sigma)$$

точно таким же образом, что и классический статистик. Но вдобавок байесовец использует и дополнительную информацию о β , хоть и нечеткую, в так называемых априориях, например,

$$\beta \sim N(h, H).$$

Можно показать, что тогда апостериорным распределением будет $p(\beta|y) \sim N(\hat{\beta}, V)$, где

$$V = (H^{-1} + X'\Sigma^{-1}X)^{-1}, \quad \hat{\beta} = V(H^{-1}h + X'\Sigma^{-1}y).$$

Среднее апостериорного распределения $\hat{\beta}$ можно рассматривать как «оценку» β , и видно, что это матрично-взвешенное среднее между априорным средним h и классической ОМНК-оценкой $(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y$. Байесовец предпочитает говорить о точности, а не о дисперсии, что есть обратная величина к точности. Тогда $V^{-1} = H^{-1} + X'\Sigma^{-1}X$, или словесно

$$\text{апостериорная точность} = \text{априорная точность} + \text{точность данных}.$$

Таким образом, точность всегда увеличивается (дисперсия уменьшается), когда информация добавляется, будь то в форме данных или априорий. Если $H^{-1} = 0$, то априорной информации нет, и классические результаты возникают как частных случаев.

В рамках нормальной байесовской модели, т.е. где и правдоподобие, и априории основаны на нормальном распределении, апостериорное распределение также нормальное (как мы только что видели), и поэтому нет математической разницы между данными и априориями, хотя, конечно же, концептуальная разница есть. Это простое наблюдение приводит к дальнейшим эквивалентностям, которые исследуются в этой заметке.

Причина исследовать эти эквивалентности – одновременно и концептуальная, и вычислительная. В байесовской постановке раздела 2 формулы не поддаются вычислениям, когда проектная матрица X имеет «большую» размерность $n \times k$ и является «разреженной». Матрица «разрежена», когда в ней много структурных нулей. Если матрица размерности $n \times k$ имеет s структурных нулей, то эту матрицу можно сохранить как матрицу размерности $(nk - s) \times 3$, где i -я строка содержит индекс строки, индекс столбца и i -е ненулевое значение. Так действовать полезно в случае, когда память для хранения важнее, чем скорость доступа. С недавних пор большие разреженные матрицы стали важны для экономистов, например, при изучении панелей в более чем миллион работающих в более чем 500.000 фирмах; см. Abowd, Kramarz & Margolis (1999) и Abowd, Creedy & Kramarz (2002), использующие алгоритм Dongarra, Duff, Sorenson & van der Vorst (1991), или при оценивании системы национальных счетов, особенно в развивающихся странах, при котором можно встретить 5.000 и более переменных и 20.000 наблюдений; см. Magnus, van Tongeren & de Vos (2000), использующие программное обеспечение, разработанное в Danilov & Magnus (2007).

В разделе 2 формулируется основной вопрос, выраженный в байесовской терминологии. В разделе 3 показывается эквивалентность байесовской задачи и различных оптимизационных задач, включая наименьшие квадраты с ограничениями и без и наилучшее линейное несмещенное оценивание. Все эти методы эквивалентны задаче наименьших квадратов без ограничений. В разделе 4 результаты кратко интерпретируются. Наконец, в разделе 5 содержатся некоторые мысли относительно доказательства матричных равенств.

2 Данные и априории

Нам интересен вектор $\beta = (\beta_1, \beta_2, \dots, \beta_k)'$, состоящий из k латентных случайных величин. Доступны данные по n_1 линейным комбинациям β . Пусть y_1 обозначает вектор данных размерности $n_1 \times 1$. Наша отправная точка – уравнение наблюдения

$$y_1 | \beta \sim N_{n_1}(X_1 \beta, \Sigma_1). \quad (1)$$

Матрица X_1 размерности $n_1 \times k$ часто принимает форму отборочной матрицы, скажем, $X_1 = (I_{n_1}, 0)$, так что $X_1 \beta$ – подвектор β , хотя это не является необходимым. Также не требуется, чтобы матрица X_1 имела полный ранг.¹ Наблюдения несмещены в том смысле, что $\mathbb{E}(y_1 | \beta) = X_1 \beta$. Матрица Σ_1 размерности $n_1 \times n_1$ обозначает положительно определенную дисперсионную матрицу, обычно (хотя необязательно) диагональную.

Вдобавок к n_1 данным у нас есть доступ к еще двум единицам информации – это априорные соображения, касающиеся латентных переменных или их линейных комбинаций, и детерминистические линейные ограничения. В частности, у нас m_1 случайных априорий:

$$R_1 \beta \sim N_{m_1}(h_1, H_1) \quad (2)$$

и m_2 точных ограничений (тождеств):

$$R_2 \beta = h_2 \text{ (почти наверное),} \quad (3)$$

что составляет $m := m_1 + m_2$ единиц априорной информации.

Мы предполагаем, что матрица H_1 размерности $m_1 \times m_1$ положительно определена (и поэтому невырождена), и что матрица R_2 размерности $m_2 \times m_1$ имеет полный строковый ранг m_2 (так что точные ограничения линейно независимы и поэтому образуют совместную систему уравнений). Определим

$$R := \begin{pmatrix} R_1 \\ R_2 \end{pmatrix}, \quad h := \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}, \quad H := \begin{pmatrix} H_1 & 0 \\ 0 & 0 \end{pmatrix},$$

¹Условие полного ранга для X_1 сделано в Magnus, van Tongeren & de Vos (2000, теорема 1), но на самом деле оно излишне.

и предположим, что $\text{rk}(R) = m$, что, естественно, означает, что и R_1 , и R_2 имеют полный строковый ранг. Позже мы увидим, что ранговое условие на R не является серьезным ограничением, поскольку нет математической разницы в случае нормальности между априориями и данными, так что всегда можно рассматривать все априории как данные и наоборот. Поэтому условие $m \leq k$ также неограничительно.

Чтобы идентифицировать все k переменных из информации (данных и априориях), необходимы по крайней мере k единиц информации: $m + n_1 \geq k$. Но этого недостаточно для идентификации, ибо частично информация может приходиться на одни и те же переменные. Если $m = k$, все переменные идентифицированы. Если $m < k$, вводится полуортогональная матрица L размерности $k \times (k - m)$, такая, что $RL = 0$ и $L'L = I_{k-m}$, и тогда необходимым и достаточным для идентифицируемости является условие

$$\text{rk}(X_1 L) = k - m.$$

Другим эквивалентным условием является

$$\text{rk} \begin{pmatrix} R \\ X_1 \end{pmatrix} = k.$$

Это следует из того, что определение L влечет за собой

$$\text{rk} \begin{pmatrix} R \\ X_1 \end{pmatrix} = \text{rk}(R) + \text{rk}(X_1 L).$$

3 Эквивалентности

Мы представим шесть оценок β (и их дисперсии), все шесть эквивалентны. Эта эквивалентность основана на двух фактах. Во-первых, байесовский анализ с нормально распределенными данными и нормальными априориями тесно связан с квадратичной задачей минимизации. Во-вторых, наилучшее линейное несмещенное оценивание также тесно связано с квадратичной минимизацией (наименьшими квадратами).

Байесовское решение

Используя теорему 1 в Magnus, van Tongeren & de Vos (2000), видим, что апостериорное распределение β есть

$$\beta|y_1 \sim N_k(\hat{\beta}, V),$$

где

$$V = R^+ H R^{+'} - R^+ H R^{+'} X_1' \Sigma_0^{-1} X_1 R^+ H R^{+'} + C K C' \quad (4)$$

и

$$\hat{\beta} = R^+ h - (R^+ H R^{+'} + C K) X_1' \Sigma_0^{-1} (X_1 R^+ h - y_1). \quad (5)$$

В этих выражениях задействованы следующие обозначения:

$$\Sigma_0 := \Sigma_1 + X_1 R^+ H R^{+'} X_1', \quad C := I_k - R^+ H R^{+'} X_1' \Sigma_0^{-1} X_1$$

и

$$K := \begin{cases} L(L' X_1' \Sigma_0^{-1} X_1 L)^{-1} L' & \text{если } m < k, \\ 0 & \text{если } m = k. \end{cases}$$

Кроме того, A^+ здесь обозначает обратную матрицу Мура–Пенроуза для матрицы A . Хотя результат выше интересен теоретически, на практике его использовать неразумно, особенно при больших размерностях. Поэтому мы ищем альтернативные эквивалентные формулировки для этих апостериорных моментов.

Только данные, никаких случайных априорий

Первый шаг к упрощению – интерпретировать все случайные априории как данные и рассматривать новый вектор «данных» размерности $n := n_1 + m_1$

$$y := \begin{pmatrix} y_1 \\ h_1 \end{pmatrix}.$$

Введя

$$X := \begin{pmatrix} X_1 \\ R_1 \end{pmatrix}, \quad \Sigma := \begin{pmatrix} \Sigma_1 & 0 \\ 0 & H_1 \end{pmatrix},$$

можно переписать уравнение измерения в виде

$$y|\beta \sim N_n(X\beta, \Sigma)$$

вместе с априориями (точными ограничениями)

$$R_2\beta = h_2 \text{ (почти навверное).}$$

Предполагая, что $m_2 < k$, введем полуортогональную матрицу L_2 размерности $k \times (k - m_2)$, такую, что $R_2L_2 = 0$ и $L_2'L_2 = I_{k-m_2}$. Условием идентифицируемости тогда будет

$$\text{rk} \begin{pmatrix} R_2 \\ X \end{pmatrix} = k,$$

или же

$$\text{rk}(XL_2) = k - m_2.$$

Апостериорные моменты $\beta|y$ в этом случае равны

$$\hat{\beta} = R_2^+ h_2 - V X' \Sigma^{-1} (X R_2^+ h_2 - y) \quad (6)$$

и

$$V = L_2 (L_2' X' \Sigma^{-1} X L_2)^{-1} L_2'. \quad (7)$$

Эти два момента численно идентичны представленным в (5) и (4).

Наилучшее линейное несмещенное оценивание

Альтернативный взгляд, также приводящий к тем же результатам, рассматривает регрессионную модель

$$y \sim N(X\beta, \Sigma)$$

с учетом линейных ограничений

$$R_2\beta = h_2,$$

где β теперь – *неслучайный* вектор параметров, которые надо оценить. Наилучшая линейная несмещенная оценка β строится как

$$\hat{\beta} = G^{-1} X' \Sigma^{-1} y + G^{-1} R_2' (R_2 G^{-1} R_2')^{-1} (h_2 - R_2 G^{-1} X' \Sigma^{-1} y) \quad (8)$$

и имеет дисперсию

$$V = G^{-1} - G^{-1} R_2' (R_2 G^{-1} R_2')^{-1} R_2 G^{-1}, \quad (9)$$

где $G := X' \Sigma^{-1} X + R_2' R_2$; см. Магнус & Нейдекер (2002, теорема 13.6). Вновь выражения (8) и (9) численно идентичны тем, что в (5) и (4).

Наименьшие квадраты при ограничениях

Существует тесная взаимосвязь между наилучшим линейным несмещенным оцениванием и наименьшими квадратами (Rao, 1971, 1973). Как показано в Магнус & Нейдекер (2002, теорема 13.16), можно получить $\hat{\beta}$ также как решение задачи

$$\begin{aligned} \min \quad & (y - X\beta)' \Sigma^{-1} (y - X\beta) \\ \text{s.t.} \quad & R_2 \beta = h_2. \end{aligned}$$

Как обсуждалось во введении, наименьшие квадраты – детерминистический метод. Все задачи наилучшего линейного несмещенного оценивания допускают эквивалентные формулировки в терминах наименьших квадратов, но взвешивающая матрица в каждом случае разная. Конечно же, эквивалентность имеет место, только если используется верная взвешивающая матрица, в данном случае Σ^{-1} .

Наименьшие квадраты без ограничений 1

Можно пойти другим путем и рассматривать ограничение как часть данных (с нулевой дисперсией). Тогда необходимо искать решение задачи

$$\min \begin{pmatrix} y - X\beta \\ h_2 - R_2\beta \end{pmatrix}' \begin{pmatrix} \Sigma + XX' & XR_2' \\ R_2X' & R_2R_2' \end{pmatrix}^{-1} \begin{pmatrix} y - X\beta \\ h_2 - R_2\beta \end{pmatrix},$$

как показано в следствии 1 из теоремы 13.15 в Магнус & Нейдекер (2002).

Наименьшие квадраты без ограничений 2

Для получения последней эквивалентности нужно явно вывести ограничения, таким образом уменьшая размерность задачи. Это достигается записью

$$R_2 = (R_{21} : R_{22}),$$

где R_{21} – матрица размерности $m_2 \times (k - m_2)$, а R_{22} – невырожденная матрица размерности $m_2 \times m_2$. После соответствующей разбивки вектора β ограничение записывается как

$$R_{21}\beta_1 + R_{22}\beta_2 = h_2,$$

так что

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} I \\ -R_{22}^{-1}R_{21} \end{pmatrix} \beta_1 + \begin{pmatrix} 0 \\ R_{22}^{-1}h_2 \end{pmatrix} \equiv Q\beta_1 + q.$$

Пусть

$$X^* := \Sigma^{-1/2}X = \begin{pmatrix} \Sigma_1^{-1/2}X_1 \\ H_1^{-1/2}R_1 \end{pmatrix}, \quad y^* := \Sigma^{-1/2}y = \begin{pmatrix} \Sigma_1^{-1/2}y_1 \\ H_1^{-1/2}h_1 \end{pmatrix}.$$

Тогда задачу с ограничениями

$$\begin{aligned} \min \quad & (y - X\beta)' \Sigma^{-1} (y - X\beta) \\ \text{s.t.} \quad & R_2 \beta = h_2. \end{aligned}$$

можно переписать как

$$\min \|(y^* - X^*q) - X^*Q\beta_1\|^2$$

по отношению к β_1 . Это простая задача наименьших квадратов без ограничений.

4 Интерпретация

Цель этой заметки в том, чтобы продемонстрировать эквивалентность трех методов при нормальности: наименьшие квадраты, наилучшее линейное несмещенное оценивание и байесовское оценивание. К доказанным эквивалентностям интерес не только концептуальный, они важны и с практических позиций. В приложениях современной эконометрики размерности матриц могут быть очень большими, например, при работе с финансовыми данными или с данными национальных счетов. Стандартные операции типа обращения (или еще хуже, обращения Мура–Пенроуза) могут стать неустойчивыми или даже недоступными. Использование устойчивых и простых формул тогда существенно.

Решение задачи наименьших квадратов при ограничениях с помощью двухступенчатой процедуры имеет то преимущество, что размерность системы значительно уменьшается. Кроме того, во многих практических ситуациях первый шаг может быть сделан единожды, с последующим использованием результатов редуцирования для множества задач наименьших квадратов при ограничениях. В частности, матрица R_2 обычно фиксирована, поскольку она представляет экономическую структуру, в то время как матрицы, относящиеся к R_1 – априории, которые меняются. Следовательно, метод, где вычисления, связанные с R_2 , производятся только единожды, будет иметь практическую важность.

Эквивалентности, обрисованные в этой заметке, имеют точный характер, никакая информация не теряется. Если размерности увеличиваются и далее, то даже самая простая формулировка, приведенная здесь, может стать неустойчивой. В таких случаях нужно искать «робастные» альтернативы, то есть методы, являющиеся в вычислительном отношении выполнимыми, *необязательно* точные, но не отклоняющиеся сильно от точных решений, даже в крайних ситуациях.

5 Некоторые мысли относительно доказательства матричных равенств

Предположим, мы хотим доказать, что $A = B$ для двух данных матриц A и B , таких как (4) и (7), или (5) и (6). Какие методы нам доступны?

Во-первых, можно просто попробовать доказать напрямую, что $A = B$, например, доказав, что $a_{ij} = b_{ij}$ для всех i и j . Этот метод обычно плох. Второй метод, несколько лучше, рассматривает $\Delta := A - B$ и доказывает $\Delta = 0$. В-третьих, что обычно быстрее, рассматривается не матричное уравнение $\Delta = 0$, а векторное уравнение $\Delta x = 0$ для всех векторов x . Или, что то же самое, можно попробовать доказать скалярное уравнение $x' \Delta' \Delta x = 0$ для всех x . Этот третий метод в сущности геометрический: мы рассматриваем отображения из x в Ax и Bx . Если результат этих двух отображений один и тот же для каждого x , то и сами отображения должны быть равны.

Четвертую и несколько отличающуюся от этих идею можно использовать, если Δ зависит от матрицы X , так что необходимо доказать, что $\Delta(X) = 0$ для каждого X . Тогда, обозначив за d дифференциал, достаточно доказать, что

$$d(\Delta(X)) = 0, \quad \Delta(X_0) = 0$$

для некоторой произвольной подобающим образом выбранной матрицы X_0 (обычно это нулевая или единичная матрица); см. упражнение 13.69 в Abadir & Magnus (2005) в качестве примера.

Интересно, что равенства в этой заметке не доказывались ни одним из этих методов. Вместо этого мы полагались на тот факт, что структуры, в рамках которых возникают уравнения, являются эквивалентными, и, следовательно, выводимые выражения также должны совпадать. Доказать равенства напрямую, конечно, возможно, но очень трудоемко.

Благодарности

Авторы благодарят редактора журнала за конструктивные замечания.

Список литературы

- Магнус, Я.Р., П.К. Катышев & А.А. Пересецкий (2005). *Эконометрика: Начальный курс*, 7-е издание. Москва: Дело.
- Магнус Я.Р. & Х. Нейдекер (2002). *Матричное дифференциальное исчисление с приложениями к статистике и эконометрике*. Москва: Физматлит.
- Abadir, K.M. & J.R. Magnus (2005). *Matrix Algebra. Econometric Exercises*, volume 1. New York: Cambridge University Press.
- Abowd, J.M., R.H. Creedy & F. Kramarz (2002). Computing person and firm effects using linked longitudinal employer-employee data. Working Paper, Cornell University.
- Abowd, J.M., F. Kramarz & D.N. Margolis (1999). High wage workers and high wage firms. *Econometrica* 67, 251–333.
- Danilov, D. & J.R. Magnus (2007). On the estimation of a large sparse Bayesian system: The Snaer project. Working Paper, Tilburg University.
- Dongarra, J.J., I.S. Duff, D.C. Sorenson & H.A. van der Vorst (1991). *Solving Linear Systems on Vector and Shared Memory Computers*. Philadelphia: SIAM.
- Magnus, J.R., J.W. van Tongeren & A.F. de Vos (2000). National account estimation using indicator ratios. *The Review of Income and Wealth* 46, 329–350.
- Rao, C.R. (1971). Unified theory of linear estimation. *Sankhyā A* 33, 371–477. Corrigenda. *Sankhyā A* 34, 477.
- Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*, 2nd edition. New York: John Wiley.

Some equivalences in linear estimation

Dmitry Danilov

Eindhoven University of Technology, Netherlands

Jan R. Magnus

Tilburg University, Netherlands

Under normality, the Bayesian estimation problem, the best linear unbiased estimation problem, and the restricted least-squares problem are all equivalent. As a result we need not compute pseudo-inverses and other complicated functions, which will be impossible for large sparse systems. Instead, by reorganizing the inputs, we can rewrite the system as a new but equivalent system which can be solved by ordinary least-squares methods.

Keywords: Linear Bayes estimation, best linear unbiased, least squares, sparse problems, large-scale optimization

Классификация JEL: C11, C61, C63