Меченов А.С.

Метод наименьших расстояний для конфлюентной модели с гомоскедастичными измерениями столбцов матрицы и правой части.

Автором в работах [1,2,3] в самых общих предположениях проведено доказательство метода наименьших расстояний оценивания параметров конфлюентных моделей пассивного эксперимента. В данной работе подробно исследуется, возможно, наиболее распространенный частный случай такой модели с различной по векторам измерений, но гомоскедастичной внутри каждого вектора, дисперсией. Результат иллюстрируется примером из эконометрии.

1. Основная модель.

Рассмотрим основное предположение [2,3] для наиболее типичного случая, когда все ошибки в столбцах конфлюентной матрицы **X** и в векторе отклика **y** гомоскедастичны.

Предположение 1. *Будем исходить из следующей линейной стохастической модели пассивного эксперимента*:

$$\begin{cases} \mathbf{y} = \Xi\beta + \mathrm{H}\delta + \mathbf{e},\ \mathbf{y} \in \mathrm{R}^n,\ \beta \in \mathrm{R}^m, \mathrm{H} \in \mathrm{R}^{nk}, \delta \in \mathrm{R}^k, E\mathbf{e} = \mathbf{0},\ E\mathbf{e}\mathbf{e}^T = \sigma^2\mathrm{I}, \\ \mathbf{X} = \Xi + \mathbf{C},\ \mathbf{X} \in \mathrm{R}^{nm},\ E\mathbf{C} = \mathbf{0}, E\mathbf{e}\underline{\mathbf{C}}^T = \mathbf{0},\ E\mathbf{c}_i\mathbf{c}_j^T = \delta_{ij}\,\mu_i^2\mathrm{I}, i,j = \overline{1,m}. \end{cases} \tag{1}$$

с матрицей $[\Xi, \mathrm{H}]$ *линейного функционального уравнения*

$$\vartheta = \Xi\beta + \mathrm{H}\delta, \vartheta \in \mathrm{R}^n, \Xi \in \mathrm{R}^{nm}, \mathrm{H} \in \mathrm{R}^{nk}, \beta \in \mathrm{R}^m, \delta \in \mathrm{R}^k, \tag{2}$$

полного ранга (δ_{ij} *- символ Кронекера*).

То есть в модели (1) пассивные наблюдения, которые будем считать нормально распределенными, имеют одинаковую дисперсию σ^2 в **y** и одинаковые дисперсии μ_i^2 в каждом i-том столбце матрицы **X**. Тогда задача оценивания параметров будет иметь вид:

Задача 1. *Зная одну реализацию* $\widetilde{\mathbf{y}}$ *и* $\widetilde{\mathbf{X}}$ $(rank\,\widetilde{\mathbf{X}} = m)$:

$$\begin{cases} \widetilde{\mathbf{y}} = \Xi\beta + \mathrm{H}\delta + \widetilde{\mathbf{e}}, \\ \widetilde{\mathbf{X}} = \Xi + \widetilde{\mathbf{C}}, \end{cases}$$

случайных величин **y** *и* **X**, *точно теоретическую матрицу* **H**, *дисперсии* $\sigma^2, \mu_i^2, i = \overline{1,m}$; *оценить неизвестные элементы* ϑ, Ξ *и неизвестные параметры* β, δ *линейной стохастической модели* (1) *методом максимального правдоподобия.*

Т.е., учитывая в данном случае тот факт, что ошибки измерений не зависят от параметров [3]получим, что параметры надо оценить таким образом, чтобы квадрат расстояния измерений $\widetilde{\mathbf{y}}, \widetilde{\mathbf{x}}_i, i = \overline{1,m}$; от искомых значений $\vartheta, \xi_i, i = \overline{1,m}$;

$$S^2 = \sigma^{-2}(\widetilde{\mathbf{y}} - \vartheta)^T(\widetilde{\mathbf{y}} - \vartheta) + \sum_{i=1}^{m}\mu_i^{-2}(\widetilde{\mathbf{x}}_i - \xi_i)^T(\widetilde{\mathbf{x}}_i - \xi_i) \tag{3}$$

в линейных ограничениях $\vartheta = \Xi\beta + \mathrm{H}\delta$ (2) был бы минимален (иначе говоря, методом наименьших расстояний -м.н.р.).

Теорема 1. *Оценки* $\hat{\mathbf{b}}, \hat{\mathbf{d}}$ *параметров* β, δ *линейной модели* (1) *из Задачи* 1 *доставляют минимум квадратичной форме*

$$S^2 = \frac{\left\|\widetilde{\mathbf{X}}\beta + \mathrm{H}\delta - \widetilde{\mathbf{y}}\right\|^2}{\sigma^2 + \sum_{i=1}^{m}\mu_i^2\beta_i^2}, \tag{4}$$

или (как можно показать с помощью дифференцирования) вычисляются из следующей нормальной системы линейных алгебраических уравнений (с.л.а.у.):

$$\begin{cases} \left(\widetilde{\mathbf{X}}^T \widetilde{\mathbf{X}} - diag\left(\mu_1^2, \ldots, \mu_m^2\right)\hat{S}^2 \mathrm{I}\right)\hat{\mathbf{b}} + \widetilde{\mathbf{X}}^T \mathrm{H}\hat{\mathbf{d}} = \widetilde{\mathbf{X}}^T \widetilde{\mathbf{y}}, \\ \mathrm{H}^T \widetilde{\mathbf{X}}\hat{\mathbf{b}} + \mathrm{H}^T \mathrm{H}\hat{\mathbf{d}} = \mathrm{H}^T \widetilde{\mathbf{y}}, \end{cases} \tag{5}$$

где $\hat{S}^2$ - минимальное значение остаточной суммы квадратов S^2 (4).

Теорема является частным случаем доказанного в [2,3] утверждения с более общим видом ковариационной матрицы.

2. Частный случай.

В качестве следствия рассмотрим построение оценки для линейной двупараметрической стохастической конфлюентной модели с различной для каждой переменной, но гомоскедастичной внутри измерений дисперсией:

$$\begin{cases} \mathbf{y} = \beta_1 \xi_1 + \beta_2 \xi_2 + \delta + \mathbf{e},\ \mathbf{y} \in \mathrm{R}^n,\ \beta \in \mathrm{R}^1, \delta \in \mathrm{R}^1, E\mathbf{e} = 0,\ E\mathbf{e}\mathbf{e}^T = \sigma^2 \mathrm{I}, \\ \mathbf{x}_1 = \xi_1 + \mathbf{c}_1,\ \mathbf{x}_1 \in \mathrm{R}^n,\ E\mathbf{c}_1 = 0, E\mathbf{e}\mathbf{c}_1^T = 0,\ E\mathbf{c}_1\mathbf{c}_1^T = \mu_1^2 \mathrm{I}, \\ \mathbf{x}_2 = \xi_2 + \mathbf{c}_2,\ \mathbf{x}_2 \in \mathrm{R}^n,\ E\mathbf{c}_2 = 0, E\mathbf{e}\mathbf{c}_2^T = 0,\ E\mathbf{c}_1\mathbf{c}_2^T = 0, E\mathbf{c}_2\mathbf{c}_2^T = \mu_2^2 \mathrm{I}, \end{cases} \tag{6}$$

и одним свободным (регрессионным) параметром δ.

Как следует из теоремы 1, оценки $\hat{b}_1, \hat{b}_2, \hat{d}$ параметров β_1, β_2, δ линейной модели (6) из Задачи 1 доставляют минимум квадратичной форме

$$S^2 = \frac{\left\|\beta_1 \widetilde{\mathbf{x}}_1 + \beta_2 \widetilde{\mathbf{x}}_2 + \delta - \widetilde{\mathbf{y}}\right\|^2}{\sigma^2 + \mu_1^2 \beta_1^2 + \mu_2^2 \beta_2^2}, \tag{7}$$

или является решением следующей с.л.а.у.:

$$\begin{bmatrix} \sum_{i=1}^n \widetilde{x}_{1i}^2 - \mu_1^2 \hat{S}^2 & \sum_{i=1}^n \widetilde{x}_{1i}\widetilde{x}_{2i} & \sum_{i=1}^n \widetilde{x}_{1i} \\ \sum_{i=1}^n \widetilde{x}_{1i}\widetilde{x}_{2i} & \sum_{i=1}^n \widetilde{x}_{2i}^2 - \mu_2^2 \hat{S}^2 & \sum_{i=1}^n \widetilde{x}_{2i} \\ \sum_{i=1}^n \widetilde{x}_{1i} & \sum_{i=1}^n \widetilde{x}_{2i} & n \end{bmatrix} * \begin{bmatrix} b_1 \\ b_2 \\ d \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n \widetilde{x}_{1i}\widetilde{y}_i \\ \sum_{i=1}^n \widetilde{x}_{2i}\widetilde{y}_i \\ \sum_{i=1}^n \widetilde{y}_i \end{bmatrix} \tag{8}$$

где $\hat{S}^2$ - минимальное по параметрам β_1, β_2, δ значение остаточной суммы квадратов S^2 (7). Очевидно, что для решения с.л.а.у. (8) необходимо использовать итерационный процесс.

3. Пример построения оценки.

Рассмотрим пример из раздела регрессионного анализа книги [4], где цена подержанного автомобиля Ниссан оценивается в функции года производства ("возраста") ξ_1 и пробега ξ_2.

Рассмотрим сначала чистую регрессию в функции возраста. Для простой линейной модели имеем оценку метода наименьших квадратов [5] (м.н.к.) вида:

$$\mathbf{y} = \delta_1 \eta_1 + \delta + \mathbf{e} \cong \underset{(0{,}28)}{-2{,}026}\eta_1 + 19{,}547,\ \mathbf{y} \in \mathrm{R}^{11}, \hat{s}^2 = 1{,}58.$$

Далее, для вычисления м.н.р.-оценки, будем считать возраст автомобиля равномерно распределенным в течение года, откуда имеем [6] для его дисперсии величину

$$c_1 \in U_{-0{,}5;0{,}5}; Ec_1 = 0;\ \mu_1^2 = Dc_1 = Ec_1^2 - \left(Ec_1\right)^2 = 1/12.$$

и будем полагать, что равномерное распределение для 11 измерений достаточно хорошо аппроксимируется нормальным.

Ниссан	$\tilde{\mathbf{x}}_1(\eta_1)$: годы	$\tilde{\mathbf{x}}_2(\eta_2)$: мили/1000	$\mathbf{y}$: цены$/1000
1	5	57	8,5
2	4	40	10,3
3	6	77	7
4	5	60	8,2
5	5	49	8,9
6	5	47	9,8
7	6	58	6,6
8	6	39	9,5
9	2	8	16,9
10	7	69	7
11	7	89	4,8

Таблица 1. Возраст автомобиля, его пробег и объявленная стоимость. В скобках приведены обозначения для регрессионного анализа.

Воспользуемся оценкой дисперсии стоимости из м.н.к. как начальным значением для итерационного процесса вычисления дисперсии цены y. Итерируя до ожидаемой величины степеней свободы S^2=11-2=9, достаточно просто подобрать отношение дисперсий исходных данных. Откуда имеем м.н.р.-оценку

$$\mathbf{y} = \beta_1\xi_1 + \delta + \mathbf{e} \cong -2{,}10\hat{\mathbf{x}}_1 + 20{,}0,\ \hat{S}^2 = 9{,}0; \mathbf{y} \in \mathrm{R}^{11},\ \sigma^2 = 1{,}2;\ \mu_1^2 = 1/12.$$

Расчет м.н.р. точнее оценивает стоимость двухлетнего автомобиля, чем то недостаточно точное описание (как это отмечено в[4]), что получено для этого автомобиля из м.н.к. В итоге м.н.р.-оценки цен полностью согласуются с исходными данными. Из нее следует, что цена нового автомобиля $20.000 и срок его службы ~10 лет. На рис. 1 содержатся исходные данные и обе оценки.

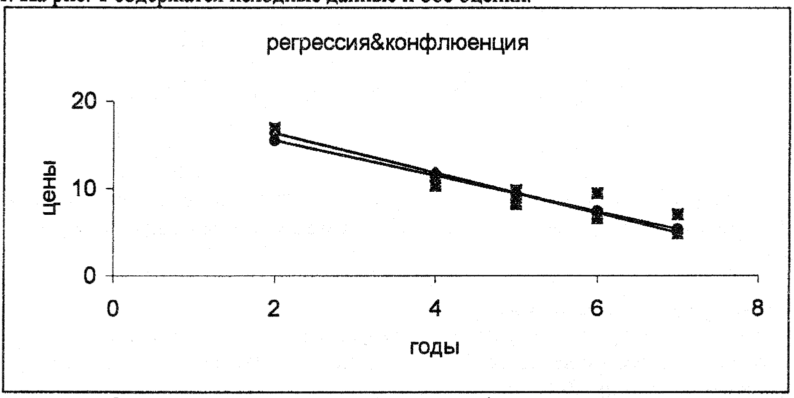

Рис. 1. Однопараметрическая модель. Квадратами обозначены исходные данные, точками - м.н.к.-оценка, ромбами - м.н.р.-оценка.

Рассмотрим теперь регрессию в функции пробега. Для простой линейной модели имеем м.н.к.-оценку

$$\mathbf{y} = \delta_2\eta_2 + \delta + \mathbf{e} \cong \underset{(0{,}016)}{-0{,}136}\eta_2 + 16{,}204, \mathbf{y} \in \mathrm{R}^{11}, \hat{s}^2 = 1{,}2; \qquad (9)$$

и м.н.р.-оценку (считаем ошибку пробега также равномерно распределенной)

$$\mathbf{y} = \beta_2\xi_2 + \delta + \mathbf{e} \cong -0{,}136\tilde{\mathbf{x}}_2 + 16{,}205,\ \hat{S}^2 = 9{,}0; \mathbf{y} \in \mathrm{R}^{11},\ \sigma^2 = 1{,}2;\ \mu_2^2 = 1/12.$$

Здесь расчет м.н.р. практически совпадает с описанием, полученным из м.н.к.-расчетов, так как теоретическая дисперсия ошибки пробега очень мала, что не совсем согласуется с исходными данными.

Замечание. Действительно, показания спидометра автомобиля напрямую зависят от размера колес, а изменение радиуса колеса на один сантиметр вследствие износа протектора влечет за собой в среднем 4-5% ошибку в измерении пробега, откуда в среднем $\mu_2^2 \approx 2{,}5$. Поэтому просчитаем результат с такой дисперсией. Здесь также по величине степеней свободы достаточно просто рассчитывается отношение дисперсий исходных данных:

$$\mathbf{y} = \beta_2\xi_2 + \delta + \mathbf{e} \cong -0{,}137\tilde{\mathbf{x}}_2 + 16{,}24,\ \hat{S}^2 = 9{,}0; \mathbf{y} \in \mathrm{R}^{11},\ \sigma^2 = 1{,}1;\ \mu_2^2 = 2{,}5.$$

Такая смена дисперсии не сильно влияет на оценку, поскольку на самом деле дисперсия должна быть много больше для большого пробега и такой подход требует расчетов с гетероскедастичной дисперсией [2,3], выходящих за рамки данной работы, однако проведя их, имеем оценку

$$\mathbf{y} = \beta_2\xi_2 + \delta + \mathbf{e} \cong -0{,}14\tilde{\mathbf{x}}_2 + 16{,}4,\ \hat{S}^2 = 9; \mathbf{y} \in \mathrm{R}^{11},\ \sigma^2 = 1{,}06;\ \mu_2^2 = \left(0{,}05\tilde{\mathbf{y}}\right)^2; (10)$$

которая, хотя и лучше согласуется с исходными данными, показывает тем не менее, что только пробег, также как и только возраст, не дает полного описания стоимости автомобиля. На рис. 2 также содержатся исходные данные и оценки (9) и (10).

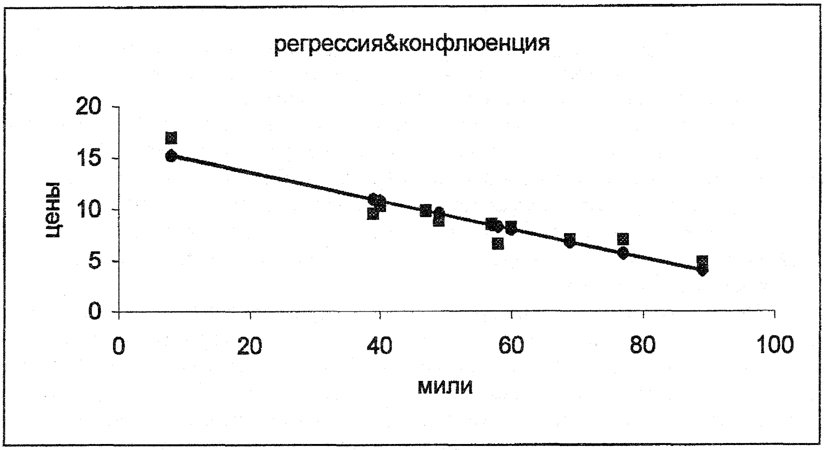

Рис. 2. Однопараметрическая модель. Квадраты - данные, ромбы - м.н.р.-оценка, точки - м.н.к.-оценка.

Рассмотрим, наконец, регрессию по обеим переменным. Для двупараметрической линейной модели имеем м.н.к.-оценку

$$\mathbf{y} = \delta_1\eta_1 + \delta_2\eta_2 + \delta + \mathbf{e} \cong -\underset{(0{,}39)}{0{,}95}\,\eta_1 - \underset{(0{,}026)}{0{,}082}\,\eta_2 + 18{,}3;\ \ \mathbf{y} \in \mathrm{R}^{11}, \hat{s}^2 = 0{,}78;$$

и м.н.р.-оценку

$$\mathbf{y} = \beta_1\xi_1 + \beta_2\xi_2 + \delta + \mathbf{e} \cong -0{,}99\tilde{\mathbf{x}}_1 - 0{,}088\tilde{\mathbf{x}}_2 + 18{,}4;\ \hat{S}^2 = 8{,}0; \mathbf{y} \in \mathrm{R}^{11}, \sigma^2 = 1;\ \mu_1^2 = 0{,}034;\ \mu_2^2 = 2{,}5;$$

И м.н.к. и м.н.р.-оценки согласуются с исходными данными. На рис. 3 также содержатся данные и обе оценки и видно, что шесть срединных точек (наиболее правильно представляющих эксперимент) лучше приближаются м.н.р.

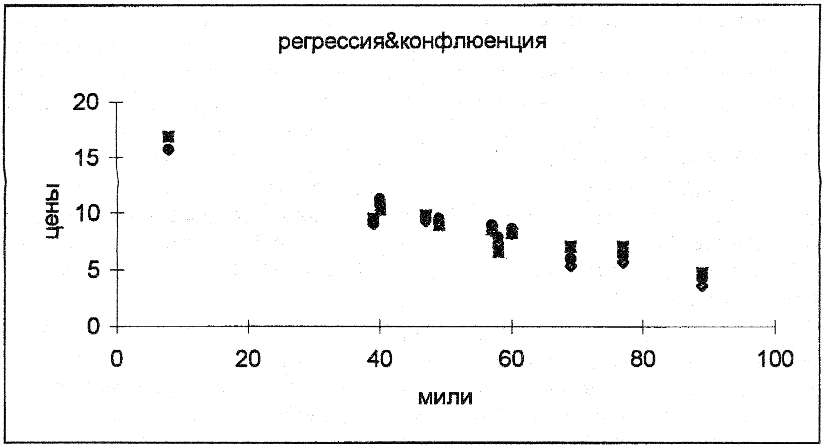

Рис. 3. Двупараметрическая линейная модель (на диаграмме в проекции плоскости регрессии на плоскость цены*пробег). Квадраты - данные, ромбы - м.н.р.-оценка, точки - м.н.к.-оценка.

Поскольку возраст и километраж сильно коррелированы (в принципе они не могут эффективно работать вместе в этой модели), то ранг матрицы (8) чувствителен к заданию величины дисперсии и м.н.р.-оценка быстро теряет устойчивость, когда дисперсии независимых переменных выравниваются. По величине степеней свободы достаточно просто рассчитывается отношение дисперсий исходных данных.

Достаточно очевидно, что модель в самом деле нелинейна по этим параметрам (поскольку цена автомобиля не равна нулю через 9-10 лет) и требует более детального рассмотрения в этом плане.

В данной работе автором подробно изучен метод построения оценок параметров для стохастических линейных конфлюентно-регрессионных моделей с различными гомоскедастичными дисперсиями - класса моделей, с одной стороны, достаточно полно описывающих события, встречающиеся в созерцательных исследованиях, с другой стороны, оценки метода нименьших расстояний для которых достаточно легко рассчитать в обычном табличном процессоре типа Excel. На приведенных примерах видно, что использование различной дисперсии может значительно улучшить простейшее описание данных задачи и позволяет полнее провести апостериорное исследование исходных данных и их причинно-следственных связей.

Литература.

1.Меченов А.С. О частично приближенных системах линейных алгебраических уравнений // Ж. вычисл. матем. и матем. физ., 1991, т. 31, N6, с. 790-799.

2.Меченов А.С. О подходе максимального правдоподобия к оценке параметров линейных функциональных соотношений. //Численные методы в математической физике. М.: Московский университет. 1996. С. 153-159.

3.Меченов А.С. О конфлюентном подходе в регрессионном анализе. // Методы математического моделирования. М.: Московский университет. 1998. С. 42-53.

4.Weiss N.A. Introductory statistics, New York. 1995.

5.Gauss K.F. Theoria motus corporum coelestium in sectioninus copecis solem ambientium, 1809,Hamburgi, (Русский перевод Н.Ф.Булаевского: Гаусс К.Ф. Избранные геодезические произведения Т.1-2. М.: Изд-во геодезической литературы. 1957. С. 152).

6.Боровков А.А. Теория вероятностей. М.: Наука, 1986, 432 с.