А.Г. Белов

О РАВЕНСТВЕ МНК И ЭЙТКЕНА ОЦЕНОК *

В статье изучается свойство равенства оценок параметров линейной множественной регрессионной модели, полученных методом наименьших квадратов (МНК) и обобщенным МНК (Эйткена) при значительно отличающихся погрешностях в исходных данных. Показан механизм формирования ошибочного суждения о качестве МНК и Эйткена оценок при визуальном графическом анализе экспериментальных данных. Проведенно численное моделирование и сравнительный анализ оценок для простой и параболической моделей.

1. Постановка задачи

Рассмотрим линейную регрессионную модель наблюдений

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

где $\mathbf{y} = (y_1, \ldots, y_n)^T \in R^n$ — вектор-столбец случайных величин (с.в.) y_i откликов, описывающих результаты i-го опыта, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T \in R^n$ — вектор-столбец случайных "ошибок" с законом распределения $\mathfrak{L}(\boldsymbol{\varepsilon})$ с математическим ожиданием (м.о.) $\mathrm{E}\,(\boldsymbol{\varepsilon}) = \underline{0}$ и известной ковариационной матрицей $\mathrm{Var}\,(\boldsymbol{\varepsilon})$; $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T \in R^k$ — вектор параметров, подлежащих оценке, $-\infty < \beta_j < +\infty$; $\mathbf{X} = \|\mathbf{x}_1, \ldots, \mathbf{x}_k\| \in R^{n\times k}$ — регрессионная матрица из вектор-столбцов (регрессоров) $\mathbf{x}_i = (x_{1i}, \ldots, x_{ni})^T$, оказывающих влияние только на среднее значение отклика $\mathrm{E}\,(y_i)$, при этом $\mathrm{rank}\,(\mathbf{X}) = k,\ k < n$. Оценку вектора параметров $\boldsymbol{\beta}$, при выполнении ряда условий, часто вычисляют с помощью метода наименьших квадратов (МНК). Как известно, если $\mathrm{Var}\,(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_n$ или не диагональная $\mathrm{Var}\,(\boldsymbol{\varepsilon}) = \mathbf{C} \in R^{n\times n}$, то решениями соответсвующих задач оценивания $\boldsymbol{\beta}$ являются МНК-оценка $\widehat{\boldsymbol{\beta}}_{ols} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ и обобщенная МНК-оценка (ОМНК или Эйткена) $\widehat{\boldsymbol{\beta}}_{gls} = (\mathbf{X}^T\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{C}^{-1}\mathbf{y}$, где $\mathbf{I}_n = \mathrm{diag}\,(1, \ldots, 1) \in R^{n\times n}$. Возникает вопрос: в каком случае, кроме частного варианта $\mathbf{C} = \sigma^2\mathbf{I}_n$, МНК и Эйткена оценки могут совпадать и на что это может влиять в практике обработки экспериментальных данных?

*Работа поддержана грантом РФФИ 14-07-00912-а.

2. Равенство оценок параметров

Т е о р е м а. *Пусть заданы с.в.* $\boldsymbol{\varepsilon}_i$, $i=1,2$ *с законом распределения* $\mathfrak{L}(\boldsymbol{\varepsilon}_i)$, *м.о.* $E(\boldsymbol{\varepsilon}_i)=\mathbf{0}$ *и ковариационной матрицей* $Var(\boldsymbol{\varepsilon}_i)=\sigma_i^2\mathbf{I}_n$. *Рассмотрим также с.в.* $\boldsymbol{\varepsilon}_3=\mathbf{P}\boldsymbol{\varepsilon}_1+\mathbf{M}\boldsymbol{\varepsilon}_2$, *где* $\mathbf{P}=\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\in R^{n\times n}$, $\mathbf{M}=\mathbf{I}_n-\mathbf{P}\in R^{n\times n}$ *— матрицы проектирования на ранговое подпространство* $R(\mathbf{X})$ *столбцов матрицы* $\mathbf{X}$ *и на ортогональное ему нуль-подпространство* $N(\mathbf{X}^T)$ *соответственно. Тогда для регрессионных моделей с векторами ошибок* $\boldsymbol{\varepsilon}_1$ *и* $\boldsymbol{\varepsilon}_3$ *соответсвующие МНК-оценка* $\widehat{\boldsymbol{\beta}}_1$ *и оценка Эйткена* $\widehat{\boldsymbol{\beta}}_3$ *совпадают.*

Д о к а з а т е л ь с т в о. Сначала определим первые два момента с.в. $\boldsymbol{\varepsilon}_3$. Для этого, воспользовавшись свойствами м.о. и дисперсии [1, с.20, 21], а также симметричностью $\mathbf{P}=\mathbf{P}^T$, $\mathbf{M}=\mathbf{M}^T$, идемпотентностью $\mathbf{P}^2=\mathbf{P}$, $\mathbf{M}^2=\mathbf{M}$ и ортогональностью $\mathbf{PM}=\mathbf{0}$ проекционных матриц $\mathbf{P}$ и $\mathbf{M}$, имеем:

$$\begin{aligned}
&\mathrm{E}(\boldsymbol{\varepsilon}_3)=\mathbf{P}\mathrm{E}(\boldsymbol{\varepsilon}_1)+\mathbf{M}\mathrm{E}(\boldsymbol{\varepsilon}_2)=\mathbf{0},\\
&\mathrm{Var}(\boldsymbol{\varepsilon}_3)=\mathbf{C}=\mathrm{Var}(\mathbf{P}\boldsymbol{\varepsilon}_1)+\mathrm{Var}(\mathbf{M}\boldsymbol{\varepsilon}_2)+2\mathrm{E}(\mathbf{P}\boldsymbol{\varepsilon}_1,\mathbf{M}\boldsymbol{\varepsilon}_2)=\\
&=\sigma_1^2\mathbf{P}+\sigma_2^2\mathbf{M}+2\mathbf{P}\mathrm{E}(\boldsymbol{\varepsilon}_1\boldsymbol{\varepsilon}_2^T)\mathbf{M}=\sigma_1^2\mathbf{P}+\sigma_2^2\mathbf{M}.
\end{aligned}$$

Докажем теперь справедливость равенств:

$$(\mathbf{X}^T\mathbf{C}^{-1}\mathbf{X})^{-1}=\sigma_1^2(\mathbf{X}^T\mathbf{X})^{-1},\ \mathbf{X}^T\mathbf{C}^{-1}=\mathbf{X}^T.$$

Действительно, для матриц $\mathbf{P}$ и $\mathbf{M}$, в силу их симметричности, справедливы [2, с.51] спектральные разложения: $\mathbf{P}=\mathbf{V}\boldsymbol{\Lambda}_P\mathbf{V}^T$, $\mathbf{M}=\mathbf{V}(\mathbf{I}_n-\boldsymbol{\Lambda}_P)\mathbf{V}^T=\mathbf{V}\boldsymbol{\Lambda}_M\mathbf{V}^T$, где $\mathbf{V}\in R^{n\times n}$ — матрица со столбцами, являющимися ортонормированными собственными векторами матрицы $\mathbf{P}$, при этом $\mathbf{V}^T\mathbf{V}=\mathbf{V}\mathbf{V}^T=\mathbf{I}_n$; $\boldsymbol{\Lambda}_P\in R^{n\times n}$ — диагональная матрица собственных значений $\mathbf{P}$, которая в силу идемпотентности $\mathbf{P}$ и ее ранга $\mathrm{rank}\,(\mathbf{P})=k$ имеет вид $\boldsymbol{\Lambda}_P=\mathrm{diag}(\underbrace{1,\ldots,1}_{k},0,\ldots,0)$, а в силу идемпотентности $\mathbf{M}$ с рангом $\mathrm{rank}\,(\mathbf{M})=n-k$ имеем $\boldsymbol{\Lambda}_M=\mathrm{diag}(0,\ldots,0,\underbrace{1,\ldots,1}_{n-k})$. Учитывая сказанное выше, справедливо представление

$$\mathbf{C}=\sigma_1^2\mathbf{P}+\sigma_2^2\mathbf{M}=\mathbf{V}\mathbf{D}\mathbf{V}^T,$$

где $\mathbf{D}=\mathrm{diag}(\underbrace{\sigma_1^2,\ldots,\sigma_1^2}_{k},\underbrace{\sigma_2^2,\ldots,\sigma_2^2}_{n-k})$. Из последнего равенства имеем:

$$\mathbf{C}^{-1}=\mathbf{V}\mathrm{diag}(\underbrace{\frac{1}{\sigma_1^2},\ldots,\frac{1}{\sigma_1^2}}_{k},\underbrace{\frac{1}{\sigma_2^2},\ldots,\frac{1}{\sigma_2^2}}_{n-k})\mathbf{V}^T=\frac{1}{\sigma_1^2}\mathbf{P}+\frac{1}{\sigma_2^2}\mathbf{M}.$$

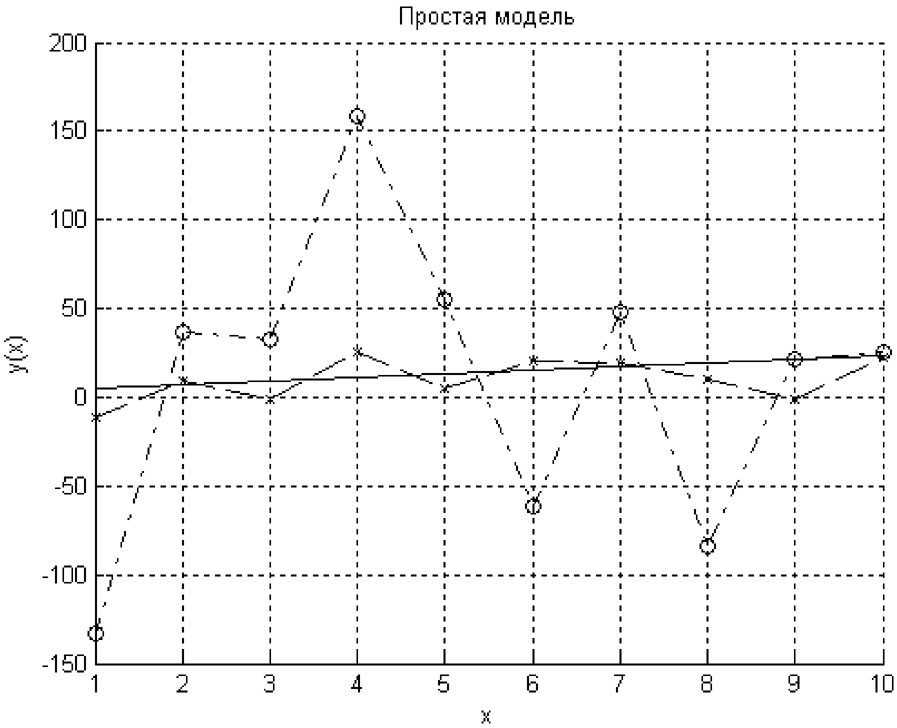

Рис. 1. Отклики: точный — сплошная линия; с ошибкой $\boldsymbol{\varepsilon}_1$ — пунктирная линия; с ошибкой $\boldsymbol{\varepsilon}_3$ — штрих-пунктирная линия

В итоге, учитывая очевидные равенства $\mathbf{PX} = \mathbf{X},\ \mathbf{MX} = \mathbf{0}$, получаем

$$(\mathbf{X}^T\mathbf{C}^{-1}\mathbf{X})^{-1} = (\frac{1}{\sigma_1^2}\mathbf{X}^T\mathbf{PX}+\frac{1}{\sigma_2^2}\mathbf{X}^T\mathbf{MX})^{-1} = \sigma_1^2(\mathbf{X}^T\mathbf{X})^{-1},\ \mathbf{X}^T\mathbf{C}^{-1} = \mathbf{X}^T.$$

З а м е ч а н и е 1. В условии теоремы нет ограничения на взаимозависимость ошибок $\boldsymbol{\varepsilon}_1$ и $\boldsymbol{\varepsilon}_2$.

З а м е ч а н и е 2. Другое доказательство теоремы может быть основано на том факте, что в ошибке $\underline{\varepsilon}_3$ вторая компонента $\mathbf{M}\boldsymbol{\varepsilon}_2$ не влияет на значение обеих оценок, поскольку принадлежит нуль-пространству $N(\mathbf{X}^T)$, а значит $\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}+\mathbf{X}^T\boldsymbol{\varepsilon}_3 = \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}+\mathbf{X}^T\mathbf{P}\boldsymbol{\varepsilon}_1+\mathbf{X}^T\mathbf{M}\boldsymbol{\varepsilon}_2 = \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}+\mathbf{X}^T\mathbf{P}\boldsymbol{\varepsilon}_1 = \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}+\mathbf{X}^T\boldsymbol{\varepsilon}_1$.

С л е д с т в и е. Из теоремы следует, что если $\mathfrak{L}(\boldsymbol{\varepsilon}_1) = N_n(\mathbf{0}, \sigma_1^2\mathbf{I}_n)$ и $\mathfrak{L}(\boldsymbol{\varepsilon}_2) = N_n(\mathbf{0}, \sigma_2^2\mathbf{I}_n)$, то, используя известное свойство нормальных с.в. [1, с.35, теорема 2.2], имеем $\mathfrak{L}(\mathbf{V}^T\boldsymbol{\varepsilon}_3) = N_n(\mathbf{0}, \mathbf{D})$. Это может быть использовано для моделирования нормальной с.в. с диагональной ковариационной матрицей и различными значениями на диагонали.

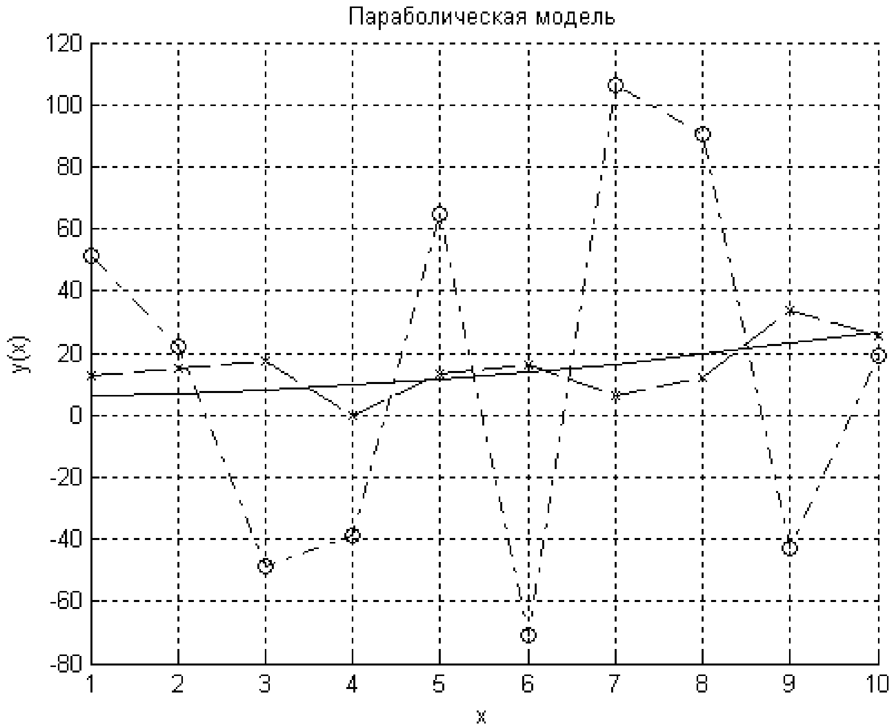

Рис. 2. Отклики: точный — сплошная линия; с ошибкой $\boldsymbol{\varepsilon}_1$ — пунктирная линия; с ошибкой $\boldsymbol{\varepsilon}_3$ — штрих-пунктирная линия

3. Численное моделирование

Проведем численные эксперименты с простой и параболической нормальными линейными регрессионными моделями, имеющими функциональные составляющие $f(x) = 1 + 2x$ и $f(x) = 6 - 3x + 2x^2$ соответственно. Для этого выберем $l = 10$ натуральных значений регрессоров. Затем для каждого из $f(x_i)$, $i = 1, \ldots, l$, независимо смоделируем $m = 1$ случайных значений y_{ij} путем аддитивного внесения в $f(x_i)$ нормальных случайных ошибок $\mathfrak{L}(\boldsymbol{\varepsilon}_1) = N_n(\mathbf{0}, \sigma_1^2 \mathbf{I}_n)$ и $\mathfrak{L}(\boldsymbol{\varepsilon}_3) = N_n(\mathbf{0}, \mathbf{C})$, при этом $\mathfrak{L}(\boldsymbol{\varepsilon}_2) = N_n(\mathbf{0}, \sigma_2^2 \mathbf{I}_n)$, $\sigma_1^2 = 100$, $\sigma_2^2 = 10000$. В результате в каждом случае получим облако из $n = lm$ значений $y_{ij} = f(x_i) + \varepsilon_{ij}$, $i = 1, \ldots, l$, $j = 1, \ldots, m$, изображенных на рис. 1, 2 и соединенных ломаными линиями: для ошибки $\boldsymbol{\varepsilon}_1$ — пунктирная, для ошибки $\boldsymbol{\varepsilon}_3$ — штрих-пунктирная. Сплошной линией изображены истинные значения наблюдаемой зависимости. Как видно, моделируемые значения y_{ij} имеют сильно различающиеся уровни погрешности. Несмотря на это решения, полученные с помощью МНК и Эйткена для соответствующих правых частей, совпадают.

4. Заключение

Таким образом, в работе показан механизм возможного образования ошибок в наблюдениях, который гарантирует неизменняемость оценки параметров МНК и Эйткена при значительно отличающихся погрешностях в исходных данных.

Проведенное численное моделирование и сравнительный анализ показали возможность ошибочного суждения о качестве получаемых МНК и Эйткена оценок по визуальному графическому анализу экспериментальных данных.

Список литературы

1. *Себер Дж.* Линейный регрессионный анализ.— М.: Мир, 1980. 456 с.
2. *Рао С. Р.* Линейные статистические методы и их применения. — М:. Наука, 1968. 548 с.