ПРИМЕНЕНИЕ ЗНАКОВЫХ КРИТЕРИЕВ К АНАЛИЗУ ЛИНЕЙНЫХ МОДЕЛЕЙ НАБЛЮДЕНИЙ

Д.Г. Белкин
Пермский государственный национальный исследовательский университет, г. Пермь

Линейные модели наблюдений получили широкое применение в разнообразных статистических приложениях. Традиционно, самым распространенных методом их анализа является метод наименьших квадратов. Этот метод позволяет получить глубокие статистические результаты в предположении, что случайные ошибки наблюдений распределены по нормальному закону. В противном случае, выводы, полученные средствами гауссовской статистики, рискуют оказаться ошибочными. Поэтому в последние десятилетия получили развитие непараметрические методы анализа линейных моделей.

К числу таких методов относится знаковый метод анализа линейных моделей [1], в основе которого лежит предположение о том, что независимые случайные ошибки наблюдений с равными вероятностями принимают положительные и отрицательные значения. При этом они даже не обязаны быть одинаково распределёнными. Так как в этом методе используются не сами наблюдения, а их знаки (или знаки остатков), естественно называть этот метод знаковым.

Важнейшим свойством знаковых процедур является независимость знаковых тестовых статистик от распределения исходных данных. Другие их привлекательные свойства – асимптотическая нормальность при минимальных ограничениях на распределение данных и устойчивость к выбросам в данных.

Общие подходы к построению знаковых процедур, ориентированных на проверку гипотез, изложены в [1]. Однако, их реализация требует значительных усилий и предполагает дополнительно глубокие знания в области линейной алгебры. Между тем применительно к анализу линейных регрессионных моделей при выполненных классических предположениях Гаусса-Маркова, включающих нормальное распределение случайных ошибок, построена простая теория [2], позволяющая проверять различные линейные гипотезы. В частности, хорошо известны t- и F- критерии для проверки гипотез о значимости отдельных коэффициентов линейной модели, о значимости модели в целом и о наличии структурных изменений в линейной модели. В данной работе приводятся результаты, существенно упрощающие реализацию знаковых критериев для решения этих трех задач.

Рассмотрим общую линейную модель наблюдений

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i,\ i = \overline{1,n} \qquad (1)$$

где $\{x_{ij}\}$ – заданные числовые значения, y_i – значения переменной y в i-м наблюдении, $\beta_0,\ldots,\beta_k$ – неизвестные коэффициенты модели, ε_i – случайная ошибка i-го наблюдения. При этом ошибки $\{\varepsilon_i, i = 1,\ldots,n\}$ являются независимыми случайными величинами, имеющими одно и то же распределение с медианой, равной нулю.

Как показано в [1], при выполнении ряда условий регулярности, накладываемых на модель (1), проверку любой линейной гипотезы о векторе параметров $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \ldots, \beta_k\right)^{\mathsf{T}}$ с помощью знаковой процедуры можно осуществить с помощью неравенств

$$\left|\mathbf{B}^T S(\hat{\mathbf{Y}})\right|^2 < q_{1-\varepsilon} \qquad (2)$$

или

$$S^{\mathsf{T}}\left(\hat{\mathbf{Y}}\right)\mathbf{B}\left(\mathbf{B}^{\mathsf{T}}\mathbf{B}\right)^{-1}\mathbf{B}^{\mathsf{T}}S\left(\hat{\mathbf{Y}}\right) < \hat{q}_{1-\varepsilon}. \qquad (3)$$

где $\mathbf{B}$ – некоторая матрица, вид которой определяется проверяемой гипотезой; $q_{1-\varepsilon}$ и $\hat{q}_{1-\varepsilon}$ – квантили распределения уровня $1-\varepsilon$ случайных величин $\left|\mathbf{B}^T \zeta\right|^2$ и $S^{\mathsf{T}}(\zeta)\mathbf{B}\left(\mathbf{B}^{\mathsf{T}}\mathbf{B}\right)^{-1}\mathbf{B}^{\mathsf{T}}S(\zeta)$ соответственно; $S(\hat{\mathbf{Y}})$ – вектор знаков оценки $\hat{\mathbf{Y}}$ вектора результатов наблюдений $\mathbf{Y} = \left(y_1,\ldots,y_n\right)^{\mathsf{T}}$, вычисляемой в условиях справедливости проверяемой гипотезы. Вектор $\boldsymbol{\zeta} = (\zeta_1,\ldots,\zeta_n)^{\mathsf{T}}$ имеет независимые компоненты с распределением: $P\left(\zeta_i = -1\right) = P\left(\zeta_i = 1\right) = 0.5,\ i = 1,\ldots,n.$

Основные результаты данной работы приведены ниже в теоремах 1 – 3. В них дано явное описание матриц $\mathbf{B}$, позволяющих построить знаковые критерии (2) – (3) для проверки гипотез о значимости отдельных коэффициентов линейной модели, о значимости модели в целом, а также о наличии структурных изменений в модели (1).

Перед тем как сформулировать теоремы, введем следующие обозначения. $\mathbf{X}_i$ – i-й столбец исходной матрицы наблюдений $\mathbf{X} = \left\| x_{i,j} \right\|, \{\mathrm{i} = \overline{1,\mathrm{n}},\ \mathrm{j} = \overline{1,\mathrm{k}}\}$. $\mathbf{V}_{(i)}$ – вектор, получаемый из вектора $\mathbf{V}$ удалением i-го элемента, $\mathbf{W}_{(i)}$ – матрица, получаемая из матрицы $\mathbf{W}$ удалением i-го столбца. $\overline{\mathrm{X}_i \mathrm{X}_j} = \frac{1}{n}\sum_{k=1}^{n} x_{i,k} x_{\mathrm{j},k}$, $\overline{\mathrm{X}_i} = \frac{1}{n}\sum_{k=1}^{n} x_{i,k}$.

Теорема 1. *В условиях регулярности проверку гипотезы о незначимости коэффициента* $H_0 : \beta_i = 0, i \in \{1, 2, ..., \mathrm{k}\}$ *в модели (1) можно осуществить с помощью матрицы* $\mathbf{B}$*, имеющей следующий вид:*

$$\mathbf{B} = -\begin{pmatrix} \mathbf{X}_1 - \overline{\mathrm{X}}_1 \cdot \mathbf{1}_n & \mathbf{X}_2 - \overline{\mathrm{X}}_2 \cdot \mathbf{1}_n & \dots & \mathbf{X}_k - \overline{\mathrm{X}}_k \cdot \mathbf{1}_n \end{pmatrix}_{(i)}^{T} \mathbf{X}_I^{-1} \mathbf{щ}_I + \left(\mathbf{X}_i - \overline{\mathrm{X}}_i \cdot \mathbf{1}_n \right),$$

где

$$\mathbf{щ}_I = \begin{pmatrix} \overline{\mathrm{X}_k} \\ \overline{\mathrm{X}_1 \mathrm{X}_k} \\ \dots \\ \overline{\mathrm{X}_k \mathrm{X}_k} \end{pmatrix}_{(i)}, \mathbf{X}_I = \begin{pmatrix} 1 & \overline{\mathrm{X}_1} & \dots & \overline{\mathrm{X}_k} \\ \overline{\mathrm{X}_1} & \overline{\mathrm{X}_1 \mathrm{X}_1} & \dots & \overline{\mathrm{X}_1 \mathrm{X}_k} \\ \dots & \dots & \dots & \dots \\ \overline{\mathrm{X}_{k-1}} & \overline{\mathrm{X}_1 \mathrm{X}_{k-1}} & \dots & \overline{\mathrm{X}_k \mathrm{X}_k} \end{pmatrix}_{(i)}$$

используя критерии (2) или (3).

Теорема 2. *Если* $rank(\boldsymbol{X}) = k + 1$, *то проверку гипотезы о незначимости модели (1) можно осуществить с помощью матрицы* $\mathbf{B}$*, имеющей следующий вид:*

$$\mathbf{B} = \begin{pmatrix} \mathbf{X}_\mathbf{1} - \overline{\mathrm{X}_1} \cdot \mathbf{1}_n & \dots & \mathbf{X}_k - \overline{\mathrm{X}_k} \cdot \mathbf{1}_n \end{pmatrix}.$$

Также, используя знаковый критерий вида (2) или (3), можно проверить гипотезу об отсутствии структурных изменений в модели (1) $H_0 : \beta_{0,1} = \beta_{0,2},\ \beta_{1,1} = \beta_{1,2}, ..., \beta_{k,1} = \beta_{k,2}$ по результатам двух серий наблюдений:

$$y_i = \beta_{01} + \beta_{11}x_{i1} + \ldots + \beta_{k1}x_{ik} + \varepsilon_i,\ i = \overline{1, n_1},$$
$$y_i = \beta_{02} + \beta_{12}x_{i1} + \ldots + \beta_{k2}x_{ik} + \varepsilon_i,\ i = \overline{n_1 + 1, n_1 + n_2}\ , n_1 + n_2 = n. \tag{4}$$

Введем дополнительно обозначения:

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_{\mathbf{I}} \\ \boldsymbol{\beta}_{\mathbf{II}} \end{pmatrix}, \qquad \boldsymbol{\beta}_{\mathbf{I}} = \left(\beta_{01}, \ldots, \beta_{k1}\right)^T, \qquad \boldsymbol{\beta}_{\mathbf{II}} = \left(\beta_{02}, \ldots, \beta_{k2}\right)^T,$$

$$\mathbf{X} = (\mathbf{X}_I \quad \mathbf{X}_{II}),\ \mathbf{X}_I = \begin{pmatrix} \mathbf{X}_I \\ \mathbf{O}_{n_2 \times (k+1)} \end{pmatrix},\ \mathbf{X}_{II} = \begin{pmatrix} \mathbf{O}_{n_1 \times (k+1)} \\ \mathbf{X}_{II} \end{pmatrix},$$

$$\mathbf{X}_I = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n_1 1} & \ldots & x_{n_1 k} \end{pmatrix},$$

$$\mathbf{X}_{II} = \begin{pmatrix} 1 & x_{n_1+1,1} & \ldots & x_{n_1+1,k} \\ 1 & x_{n_1+2,1} & \ldots & x_{n_1+2,k} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n,1} & \ldots & x_{n,k} \end{pmatrix}.$$

Здесь $\mathbf{O}_{n_i \times (k+1)}$ – нулевая матрица размерности $n_i \times \left(k+1\right)$.

Тогда справедлива

Теорема 3. *В условиях регулярности проверку гипотезы об отсутствии структурных изменений* $H_0 : \beta_{01} = \beta_{02}, \ldots, \beta_{k1} = \beta_{k2}$ *в модели (4) можно осуществить с помощью матрицы* $\mathbf{B}$*, имеющей следующий вид:*

$$\mathbf{B} = \begin{pmatrix} -\mathbf{X}_I \left(\mathbf{X}_I^T \mathbf{X}_I\right)^{-1} \mathbf{X}_{II}^T \mathbf{X}_{II} \\ \mathbf{X}_{II} \end{pmatrix}$$

используя статистики из (2) или (3).

С использованием приведенных выше теоретических результатов была проведена серия экспериментов над парной линейной моделью наблюдений. Изменяемыми параметрами являлись значения коэффициентов $\boldsymbol{\beta}$ и тип распределения случайной ошибки $\boldsymbol{\varepsilon}$. Целью экспериментов являлось сравнение мощности знаковых критериев с мощностью *t*- и *F*-критериев.

Оказалось, что при наличии в выборке выбросов применение знаковых критериев более предпочтительно. Также знаковый критерий показал лучшие результаты и при анализе выборок без выбросов при распределении случайных ошибок по закону Коши.

Список литературы

1. *Болдин М.В., Симонова Г.И., Тюрин Ю.Н.* Знаковый статистический анализ линейных моделей. М.: Наука. Физматлит, 1997. 288 с.
2. *Себер Дж.* Линейный регрессионный анализ: Пер. с англ. М.: Мир, 1980. 456 с.
3. *Шевцов Г.С.* Линейная алгебра: теория и прикладные аспекты: Учеб. Пособие. М.: Финансы и статистика, 2003. 576 с.