Тр. V международной конференции "Актуальные проблемы электронного приборостроения" АПЭП-2000. Новосибирск, 2000. - Т. 7. - С. 184-187.

ИССЛЕДОВАНИЕ РАСПРЕДЕЛЕНИЙ СТАТИСТИК КОРРЕЛЯЦИОННОГО АНАЛИЗА ПРИ ОТКЛОНЕНИИ МНОГОМЕРНОГО ЗАКОНА ОТ НОРМАЛЬНОГО[1]

Б.Ю. Лемешко, С.С. Помадин

Новосибирский государственный технический университет
Новосибирск, Россия. E-mail: headrd@fpm.ami.nstu.ru, ser@vampire.ami.nstu.ru

Аннотация. Методами статистического моделирования исследуются распределения статистик корреляционного анализа при проверке различных гипотез. В зависимости от объема выборки n исследуется сходимость распределений статистик к предельным в случае многомерного нормального закона. Исследуются распределения статистик при отличии наблюдаемого закона от многомерного нормального.

В различных приложениях статистического анализа многомерных случайных величин одну из ключевых позиций занимают задачи корреляционного анализа. В процессе решения этих задач вычисляются оценки коэффициентов и матриц парной, частной и множественной корреляции, проверяются различные статистические гипотезы относительно параметров многомерного распределения и коэффициентов корреляции. На основании результатов корреляционного анализа может делаться вывод или о наличии и характере функциональной зависимости, или о предпочтительности для описания исследуемого объекта регрессионной модели того или иного вида.

В основе существующего аппарата классического корреляционного анализа лежит предположение о принадлежности наблюдаемого случайного вектора *многомерному нормальному* закону. Базируясь на этом, получены предельные распределения статистик, используемых в классическом корреляционном анализе [1-3].

Пусть $X_1, X_2, ..., X_n$ – m-мерная выборка случайной величины объема n; M – вектор математического ожидания; Σ – ковариационная матрица с элементами σ_{ij}; $\hat{M}$ и $\hat{\Sigma}$ – оценки максимального правдоподобия (ОМП) для вектора математического ожидания и ковариационной матрицы:

$$\hat{M} = \frac{1}{n}\sum_{i=1}^{n} X_i , \quad \hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \hat{M}\right)\left(X_i - \hat{M}\right)^T .$$

При проведении данной работы предполагалось решить следующие задачи:

1. Исследовать, как быстро сходятся распределения статистик корреляционного анализа, полученные по выборкам конечного объема многомерной нормальной случайной величины, к соответствующим предельным распределениям (для тех статистик, предельные распределения которых явно не зависят от объема выборки).
2. Исследовать, что будет происходить с распределениями используемых в корреляционном анализе статистик, если наблюдаемый закон отличается от многомерного нормального.
3. Исследовать, насколько будут (или могут быть) справедливы выводы о наличии и характере функциональной зависимости или о регрессионной модели исследуемого объекта, основанные на решении задач классического корреляционного анализа, если наблюдаемый многомерный закон в большей или меньшей степени отличается от нормального.

Очевидно, что ответить на эти вопросы, используя аналитические методы, чрезвычайно сложно из-за нетривиальности этих задач. Поэтому в основу проводимого исследования положена развиваемая методика компьютерного анализа статистических закономерностей. С ее помощью проводились исследования распределений статистик, связанных с проверкой следующих гипотез классического корреляционного анализа [1-3].

[1] Работа поддержана Российским фондом фундаментальных исследований (проект № 00-01-00913)

1. О равенстве математического ожидания некоторому известному вектору $H_0 : M = M_0$.

Здесь возможны две ситуации. В случае известной ковариационной матрицы Σ используется статистика

$$X_m^2 = n(\hat{M} - M_0)^T \Sigma^{-1} (\hat{M} - M_0), \tag{1}$$

имеющая в качестве предельного распределения $G(X_m^2 \mid H_0)$ χ_m^2-распределение, с числом степеней свободы *m*.

В случае неизвестной ковариационная матрицы Σ используется статистика

$$T^2 = \frac{n(n-m)}{m(n-1)} (\hat{M} - M_0)^T \hat{\Sigma}^{-1} (\hat{M} - M_0), \tag{2}$$

которая в качестве предельного $G(T^2 \mid H_0)$ имеет $F_{m,n-m}$-распределение Фишера, с параметрами *m* и *n-m*.

2. О коэффициенте парной корреляции может проверяться гипотеза $H_0 : r_{ij} = 0$. В этом случае статистика

$$t = \frac{\sqrt{n-2}\left|\hat{r}_{ij}\right|}{\sqrt{1-\hat{r}_{ij}^2}}, \tag{3}$$

где $\hat{r}_{ij} = \dfrac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}}$, имеет в качестве предельного распределения $G(t \mid H_0)$ t_{n-2}-распределение Стьюдента, с числом степеней свободы *n*-2.

При проверке гипотезы вида $H_0 : r_{ij} = r_0$ используется статистика

$$z_0 = \sqrt{n-3}\left(\frac{1}{2}\ln\left(\frac{1+\hat{r}_{ij}}{1-\hat{r}_{ij}}\right) - \frac{1}{2}\ln\left(\frac{1+r_0}{1-r_0}\right) - \left(\frac{r_0}{2(n-1)}\right)\right), \tag{4}$$

имеющая в качестве предельного $G(z_0 \mid H_0)$ стандартное нормальное распределение $N(0,1)$.

3. Оценка частного коэффициента корреляции вычисляется по формуле

$$\hat{r}_{ij;l+1,\ldots,m} = \frac{\hat{\sigma}_{ij;l+1,\ldots,m}}{\sqrt{\hat{\sigma}_{ii;l+1,\ldots,m}\hat{\sigma}_{jj;l+1,\ldots,m}}},$$

где *l* – число компонент в условном распределении ($2 \le l \le m$);

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \quad \Sigma_{11} = \left[\sigma_{ij}\right]_{i,j=1}^{l}, \quad \Sigma_{12} = \Sigma_{21}^T = \left[\sigma_{ij}\right]_{i=1,j=l+1}^{l,m}, \quad \Sigma_{22} = \left[\sigma_{ij}\right]_{i,j=l+1}^{m},$$

$$\Sigma_{12\bullet 2} = \left[\sigma_{ij;l+1,\ldots,m}\right]_{i,j=1}^{l} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

При проверке гипотезы $H_0 : r_{ij;l+1,\ldots,m} = 0$ вычисляется статистика

$$t = \frac{\sqrt{n-m+l-2}\left|\hat{r}_{ij\bullet l+1,\ldots,m}\right|}{\sqrt{1-\hat{r}_{ij\bullet l+1,\ldots,m}}}, \tag{5}$$

которая имеет в качестве предельного распределения $G(t \mid H_0)$ $t_{n-m+l-2}$- распределение Стьюдента, с числом степеней свободы *n-m+l*-2.

При проверке гипотезы вида $H_0 : r_{ij;l+1,\ldots,m} = r_0$ используется статистика

$$z_0 = \sqrt{n-3}\left(\frac{1}{2}\ln\left(\frac{1+\hat{r}_{ij\bullet l+1,\ldots,m}}{1-\hat{r}_{ij\bullet l+1,\ldots,m}}\right) - \frac{1}{2}\ln\left(\frac{1+r_0}{1-r_0}\right) - \left(\frac{r_0}{2(n-1)}\right)\right). \qquad (6)$$

Предельным распределением $G(z_0 \mid H_0)$ этой статистики является стандартное нормальное распределение $N(0,1)$.

4. Оценка множественного коэффициента корреляции вычисляется по формуле

$$\hat{r}_{i;l+1,\ldots,m} = \sqrt{\frac{\hat{\sigma}_{(i)}\Sigma_{22}^{-1}\hat{\sigma}_{(i)}^T}{\hat{\sigma}_{ii}}},$$

где $\sigma_{(i)}$ – *i-я* строка матрицы Σ_{12}, σ_{ii} – элемент матрицы Σ_{11}.

При проверке гипотезы вида $H_0 : r_{i\bullet l+1,\ldots,m} = 0$ вычисляется статистика

$$F = \frac{n-m+l-1}{m-l}\frac{\hat{r}^2_{i\bullet l+1,\ldots,m}}{1-\hat{r}^2_{i\bullet l+1,\ldots,m}}, \qquad (7)$$

имеющая в качестве предельного распределения $G(F \mid H_0)$ $F_{m-l,n-m+l-1}$-распределение Фишера, с параметрами *m-l* и *n-m+l-1*.

Подчеркнем, что *все упомянутые предельные распределения рассмотренных статистик имеют место при наблюдении многомерного нормального закона*. Что произойдет с предельными распределениями этих статистик, насколько могут быть справедливы выводы, формулируемые на основании решения задач корреляционного анализа, если наблюдаемый закон отличается от многомерного нормального, заранее сказать нельзя. Не найдено и указаний на решение данных задач в литературных источниках.

Одной из возникающих проблем на пути исследования методами статистического моделирования предельных распределений статистик, является задача моделирования псевдослучайных векторов, “заданным образом” отличающихся от многомерного нормального.

В данной работе для моделирования многомерных распределений таких, как логистическое и Лапласа, использовался подход принятый для нормального закона [4].

Статистический анализ моделируемых псевдослучайных векторов показал, что маргинальные функции распределения получаемых псевдослучайных величин имеют хорошее согласие с одноименными одномерными законами. В то же время, пока нельзя считать, что нами полностью решена задача моделирования псевдослучайных векторов с законом распределения, “заданным образом” отличающимся от многомерного нормального.

Исследование распределений классических статистик (1)-(7) в случае многомерного нормального закона показало, что полученные эмпирические распределения статистик хорошо согласуются с предельными. Распределение статистики (1) X_m^2 хорошо согласуется с предельным, начиная с объемов выборок многомерной нормальной величины $n \geq 15$. Распределения статистик (4), (6) z_0 (парная и частная корреляции) хорошо согласуется с предельным, начиная с объемов выборок многомерной нормальной величины $n \geq 50$. Не отмечено зависимости скорости сходимости распределений статистик от размерности случайной величины *m*.

Исследование распределений статистик (1)-(7) в случае принадлежности наблюдаемого случайного вектора многомерному “логистическому” закону показало отсутствие значимых отклонений эмпирических распределений статистик от предельных распределений классических статистик. Проверка гипотез о согласии осуществлялась с использованием

критериев типа χ^2 [5] и непараметрических критериев типа Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса [6]. Полученный результат объясняется близостью многомерных нормального и логистического законов. В то же время этот результат говорит о слабой различимости данных многомерных законов с помощью статистических критериев.

В случае многомерного распределения Лапласа можно говорить о наметившемся отличии эмпирических распределений статистик от предельных распределений классических статистик. Однако в своем большинстве такие отличия оказались статистически незначимы. Исключение составляет статистика X_m^2, распределение которой существенно отклоняется от χ_m^2-распределения.

На основании проведенных исследований можно сделать следующие выводы:

1. Подтверждена эффективность применения методики компьютерного анализа статистических закономерностей для исследования распределений статистик задач корреляционного анализа многомерных случайных величин.
2. Исследована сходимость распределений ряда статистик корреляционного анализа к соответствующим предельным в зависимости от объемов выборок многомерной случайной величины.
3. Показано, что при не слишком большом отклонении многомерного закона от нормального предельные распределения большинства исследуемых статистик не претерпевают значимых изменений, за исключением статистики X_m^2, вид эмпирического распределения которой изменился существенно.
4. Результаты исследований показали возможность построения методами компьютерного анализа статистических закономерностей ***моделей*** предельных распределений статистик при любом виде наблюдаемого закона многомерной случайной величины.
5. Дальнейшей задачей является исследование ***возможности установления характера взаимозависимости*** компонент многомерного вектора при произвольном законе, а также развитие методики компьютерного анализа в процессе построения программной системы, составными частями которой предполагаются развитие систем корреляционного анализа многомерных наблюдений [7] и статистического анализа одномерных наблюдений [8] .

ЛИТЕРАТУРА

1. Андерсон Т. Введение в многомерный статистический анализ. - М.: Физматгиз, 1963. - 500 с.
2. Кендалл М., Стьюарт А. Статистические выводы и связи. – Москва: Издательство «Наука», 1973. – 900 с.
3. Кендалл М., Стьюарт А. Многомерный статистический анализ и временные ряды. – Москва: Издательство «Наука», 1976. – 736 с.
4. Ермаков С.М., Михайлов Г.А. Статистическое моделирование. – Москва: Издательство «Наука», 1982. – 296 с.
5. Денисов В.И., Лемешко Б.Ю., Постовалов С.Н. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Методические рекомендации. Часть I. Критерии типа χ^2. – Новосибирск: Издательство НГТУ, 1998. – 126 с.
6. Лемешко Б.Ю., Постовалов С.Н. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Методические рекомендации. Часть II. Непараметрические критерии. – Новосибирск: Издательство НГТУ, 1999. – 86 с.
7. Лемешко Б.Ю. Корреляционный анализ многомерных наблюдений случайных величин: Программная система. – Новосибирск: Издательство НГТУ, 1995. – 39 с.
8. Лемешко Б.Ю. Статистический анализ одномерных наблюдений случайных величин: Программная система. – Новосибирск: Издательство НГТУ, 1995. – 125 с.