ОБ ИСПОЛЬЗОВАНИИ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ ПРИ ПОСТРОЕНИИ РЕГРЕССИОННЫХ МОДЕЛЕЙ

В.В. Мокеев

Главные компоненты часто используются совместно с другими статистическими методами. Одним из таких методов является регрессионный анализ. Во множественной регрессии, одной из основных трудностей является проблема мультиколлинеарности, которая возникает, когда существует почти постоянная линейная зависимость между двумя и более эндогенными переменными. Обзор проблемы мультиколлинеарности можно найти в работе [1]. Мультиколлинеарность часто, но не всегда указывает на большую корреляцию между подмножества переменных, и, если мультиколлинеарность существует, то погрешности оценки коэффициентов регрессии могут быть очень большими, что приводит к нестабильным и потенциально обманчивым оценкам регрессионного уравнения. Для решения этой проблемы используются различные подходы.

Одно из направлений решения проблемы мультиколлинеарности базируется на использовании наборов эндогенных переменных, выбранных так, что они не содержали мультиколлинеарностей. Многочисленные методы построения таких наборов описаны в работах [2–5]. Для решения проблемы мультиколлинеарности разработан целый класс подходов, который включает ридж-регрессию, усеченные оценки, метод частично наимень-

ших квадратов, а также различные вариации регрессионного анализа на основе метода главных компонент.

Наиболее распространенный метод, известный как регрессия главных компонент, начинается с использования главных компонент эндогенных переменных вместо самих эндогенных переменных. Так как главные компоненты некоррелированны, расчеты коэффициентов регрессии могут быть сильно упрощены. Если все главные компоненты включены в регрессию, то в результате полученная регрессионная модель эквивалентна модели, полученной непосредственно по системе эндогенных переменных методом наименьших квадратов, так что погрешности, вызванные мультиколлинеарностью, при этом не исчезают. Однако, вычисления оценок методом наименьших квадратов через регрессию главных компонент могут быть численно более стабильным, чем прямые расчеты [6]. В работе показано как при построении регрессионной модели устраняется мультиколлинеарность с помощью метода главных компонент.

Построение регрессионной модели начинается с выдвижения гипотезы о том, что переменная y зависит от набора эндогенных (независимых) переменных x_i, т. е.

$$y = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n + \varepsilon \qquad (1а)$$

или в матричном виде

$$y = \mathbf{X}\alpha + \varepsilon, \qquad (1б)$$

где α – вектор неизвестных коэффициентов, y – вектор из m наблюдений зависимой переменной, измеренной относительно ее среднего значения, $\mathbf{X}$ – это матрица $(m \times n)$ независимых переменных, измеренных также относительно их средних значений, $\mathbf{X}$ представляет матрицу, каждый столбец которой содержит все значения одной независимой переменной, ε – вектор ошибок.

Регрессионный анализ проверяет статистическую состоятельность модели при данной гипотезе. Следует отметить, что регрессионный анализ не в состоянии «доказать» гипотезу, он может лишь подтвердить ее статистически или опровергнуть.

Для заданного набора данных оценки коэффициентов репрессии можно найти с помощью метода наименьших квадратов. Если оценки метода наименьших квадратов обозначить через α' то соответствующее уравнение регрессии будет иметь следующий вид:

$$\hat{y} = \mathbf{X}\alpha' + \varepsilon'.$$

Если метод наименьших квадратов используется для получения оценки для α, то записывается функционал

$$\Phi = (\mathbf{X}\ \alpha' - y)^T (\mathbf{X}\alpha' - y). \qquad (2)$$

Здесь α' – оценка вектора α. После минимизации функционала (2), получим уравнение

$$\mathbf{X}^T\mathbf{X}\alpha' - \mathbf{X}^T y = 0\,. \tag{3}$$

С учетом формулы $\mathbf{A} = \frac{1}{m}\mathbf{X}^T\mathbf{X}$ уравнение (3) преобразуется к виду

$$m\mathbf{A}\alpha' = \mathbf{X}^T y \tag{4}$$

Решая полученное уравнение, получается

$$\alpha' = {}^{1}\!/_{m}\,\mathbf{A}^{-1}\mathbf{X}^T y\,. \tag{5}$$

Решение уравнения (3) существенно упрощается, если используются собственные векторы матрицы $\mathbf{A}$. При этом матрица $\mathbf{A}$ может быть представлена в виде следующего разложения:

$$\mathbf{A} = \mathbf{V}_0\,\Sigma\mathbf{V}_0^T\,, \tag{6}$$

где $\mathbf{V}_0$ – матрица собственных векторов, $\mathbf{\Sigma}$ – диагональная матрица, диагональный элемент которой равен собственному значению λ_i матрицы $\mathbf{A}$.

Подставляя соотношение (6) в уравнение (4) получим

$$m\mathbf{V}_0\,\Sigma\mathbf{V}_0^T\,\alpha' = \mathbf{X}^T y \tag{7}$$

Далее умножаем справа на матрицу

$$m\mathbf{V}_0^T\mathbf{V}_0\,\Sigma\mathbf{V}_0^T\,\alpha' = \mathbf{V}_0^T\mathbf{X}^T y \tag{8}$$

Далее после небольших преобразований получается соотношение

$$\alpha' = {}^{1}\!/_{m}\,\mathbf{V}_0\Sigma^{-1}\mathbf{V}_0^T\mathbf{X}^T y = {}^{1}\!/_{m}\sum_{i=1}^{n}\lambda_i^{-1}\mathbf{v}_{0i}\mathbf{v}_{0i}^T\mathbf{X}^T y \tag{9}$$

Выражение (9) дает понимание того, как мультиколлинеарность приводит к большим искажениям элементов вектора α. Если мультиколлинеарность существует, то она проявляется в виде главных компонент с очень маленькими собственными значениями. Другими словами, если последние главные компоненты λ_i имеют маленькие значения, то величины $1/\lambda_i$ принимают очень большие значения. Таким образом, выражение (9) показывает, что большие искажения в коэффициентах α' связаны с присутствием в разложении (6) главных компонент, имеющих маленькие собственные значения (λ_i).

Одним из путей преодоления мультиколлинеарности является удаление членов, соответствующих очень маленьким λ_i, которое приводит вычислению оценки

$$\alpha'' = 1/m\sum_{i=1}^{p}\lambda_i^{-1}\mathbf{v}_{0i}\mathbf{v}_{0i}^T\mathbf{X}^T y \tag{10}$$

где $\lambda_{p+1}, \lambda_{p+2}, \ldots, \lambda_n$ – очень маленькие собственные значения. По сути, такой подход устранения мультиколлинеарности эквивалентен приравниванию нулю последних $(n-p)$ элементов вектора α'.

Регрессионное уравнение, использующее в качестве независимых переменных главные факторы, имеет вид

$$y = \mathbf{Z}\beta + \varepsilon. \quad (11)$$

Главные факторы определяются с помощью выражения

$$\mathbf{Z} = \mathbf{X}\mathbf{V}_0^T. \quad (12)$$

Для получения оценки вектора β также используется метод наименьших квадратов, в результате чего получаем уравнение

$$\mathbf{Z}^T\ \mathbf{Z}\beta' - \mathbf{Z}^T y = 0. \quad (13)$$

С учетом соотношения (12) уравнение (13) преобразуется к виду

$$\mathbf{V}_0\mathbf{X}^T\mathbf{X}\mathbf{V}_0^T\beta' = (\mathbf{X}\mathbf{V}_0)^T y. \quad (14)$$

В итоге, решая уравнение (14) получаем оценку вектора β

$$\beta' = 1/m\,\Sigma^{-1}\mathbf{V}_0^T\mathbf{X}^T y. \quad (15)$$

Уменьшение колебаний оценки α'', определяемой формулой (10), достигается введением смещения в оценку α'

$$\alpha'' = \alpha' - 1/m \sum_{i=m+1}^{p} \lambda_i^{-1}\mathbf{v}_{0i}\mathbf{v}_{0i}^T\mathbf{X}^T y. \quad (16)$$

Однако если мультиколлинеарность достаточно большая, то уменьшение отклонений оценки α'', существенно, в то время как само вводимое смещение сравнительно мало. Фактически, если элементы β в уравнении (11), соответствующие удаленным компонентам, реально нулевые, то смещения отсутствует.

Таким образом, на основе анализа главных компонент показан механизм возникновения искажений коэффициентов регрессионной модели, обусловленных мультиколлинеарностью независимых переменных. Показано, что путем введения сравнительно небольших смещений, можно устранить подобные искажения.

Библиографический список

1. Gunst, R.F. Regression analysis with multicollinear predictor variables: Definition, detection and effects / R.F. Gunst // Commun. Statist. Theor. Meth. –1983. – Vol. 12. – P. 2217–2260.

2. Draper, N.R. Applied Regression Analysis / N.R. Draper, H. Smith. 3rd edition. – New York: Wiley, 1998.

3. Hocking, R.R. The analysis and selection of variables in linear regression. / R.R. Hocking // Biometrics. – 1976. – Vol. 32. – P. 1–49.

4. Miller, A.J. Selection of subsets of regression variables (with discussion) /A.J. Miller // J. R. Statist. Soc. A. – 1984. – Vol. 147. – P. 389–425.

5. Miller, A.J. Subset Selection in Regression / A.J Miller // London: Chapman and Hall, 1990.

6. Flury, B. Graphical representation of multivariate data by means of asymmetrical faces / B. Flury, H. Riedwyl // J. Amer. Statist. Assoc. – 1981. – 76. – P. 757–765.