О проверке гипотез и построении прогнозов для моделей в виде систем одновременных уравнений

Борис Ю. Лемешко, Алексей Е. Щеглов
Новосибирский Государственный Технический Университет

Аннотация **– Исследовано распределение статистик критериев Андерсона-Рубина и переопределения, проанализирована процедура построения интервальных прогнозов эндогенных переменных при нарушении предположения нормальности шума и в выборках конечных размеров. Произведено сравнение точности прогнозов при использовании разных методов оценивания приведенной формы.**

Ключевые слова **– критерий Андерсона-Рубина, критерий переопределения, статистика Хотеллинга.**

I. ВВЕДЕНИЕ

МОДЕЛИ В ВИДЕ систем одновременных уравнений (СОУ) широко используются при решении различных прикладных задач, особенно при описании экономических процессов. С помощью таких моделей можно описывать поведение процентных ставок, спрос на пищевые продукты, экономику страны.

Как и для любой другой математической модели какого-либо экономического явления или процесса, основной целью построения моделей в виде СОУ является прогнозирование значений интересующих исследователя переменных. Поэтому после того, как закончено построение модели, естественным образом возникают вопросы о том, насколько удачно удалось ее специфицировать, идентифицировать и даст ли ее использование результаты, достаточно адекватные действительности. Ответы на эти вопросы может дать проверка статистических гипотез с помощью различных критериев.

Часто используемые статистические критерии основываются на предположении, что шум имеет нормальное распределение, а также их свойства носят асимптотический характер. На практике же исследователь имеет дело с ограниченными объемами наблюдений, а гипотеза о нормальности шума выполняется далеко не всегда. Если этому не придавать значения и пользоваться стандартными процедурами анализа, то можно делать ошибочные выводы.

Так, например, в [1] распределение статистики критерия Андерсона-Рубина рассматривается только при нормальном распределении шума, в [2] сказано, что статистика критерия переопределения имеет асимптотическое распределение, а при каких распределениях шума это выполняется ничего не сказано.

Процедура построения интервального прогноза эндогенных переменных, описанная в [3, 4], базируется на статистике Хотеллинга, распределение которой известно только при нормальном распределении шума. Кроме того, в [4] нет информации о распределении этой статистики в условиях малых объемов выборок, и предполагается, что среди предопределенных переменных нет лаговых эндогенных.

II. ПОСТАНОВКА ЗАДАЧИ

Таким образом, цель настоящей работы заключалась в исследовании распределения статистики Хотеллинга и статистик критериев Андерсона-Рубина и переопределения в выборках конечных размеров и при нарушении предположения нормальности шума. Отдельной задачей стояло исследование влияния на распределение статистики Хотеллинга наличия среди предопределенных переменных лаговых эндогенных. Кроме того, было проведено сравнение точности прогнозов эндогенных переменных при разных методах оценивания приведенной формы.

К решению поставленных задач возможны два подхода. Первый заключается в изучении строгой аналитической формы законов распределения статистик, а второй – в определении этих распределений эмпирически, с помощью достаточного числа искусственных выборок, полученных с использованием методики компьютерного моделирования.

За редким исключением, применение первого подхода оказывается чрезвычайно сложным из-за нетривиальности возникающих при этом задач. Тогда как второй подход представляет собой доступный и эффективный аппарат проведения таких исследований. Поэтому в работе основное внимание было уделено второму подходу.

III. ТЕОРИЯ СОУ

Пусть n – количество временных тактов, Y – матрица наблюдений за всеми эндогенными переменными размерности $n \times m$, X – матрица наблюдений за всеми предопределенными переменными размерности $n \times p$, m_i – количество эндогенных переменных, входящих в i-е уравнение СОУ, p_i – количество предопределенных переменных, входящих в i-е уравнение СОУ.

При использовании критерия Андерсона-Рубина предполагается, что i-е структурное уравнение СОУ имеет следующий вид:

$$Y^{(i)} = Y(i)\gamma\_(i) + X(i)h\_(i) + \varDelta(i), \quad (1)$$

где $Y^{(i)}$ и $Y(i)$ – матрицы эндогенных переменных размерности $n \times 1$ и $n \times m_i - 1$, соответственно, $X(i)$ – матрица предопределенных переменных размерности $n \times p_i$, $\gamma\_(i)$ и $h\_(i)$ – векторы неизвестных параметров размерности $m_i - 1$ и p_i, соответственно, $\varDelta(i)$ – вектор случайных составляющих размерности n.

Критерий Андерсона-Рубина используется для проверки гипотезы вида $H_0 : \gamma\_(i) = \gamma_0$ для уравнения (1) на основе следующей статистики [1]:

$$\frac{[SS_0 - SS_1](n-p)}{SS_1(p-p_i)}, \quad (2)$$

где $SS_0 = y^T M(X(i))y$, $SS_1 = y^T M(X)y$, $y = (Y^{(i)} - Y(i)\gamma_0)$, $P(A) = A(A^T A)^{-1} A^T$, а $M(A) = I_n - P(A)$.

При верной гипотезе статистика (2) подчиняется распределению Фишера со степенями свободы $p - p_i$ и $n - p$.

При использовании критерия переопределения предполагается, что i-е структурное уравнение СОУ имеет вид:

$$\gamma^T(i)Y_{i\_t} = h\_^T(i)X_t(i) + \varDelta_t, \quad (3)$$

где $\gamma(i)$ – вектор параметров размерности m_i, в котором i-й элемент равен 1, Y_{i_t} – вектор эндогенных переменных размерности m_i, $X_t(i)$ – вектор предопределенных переменных размерности p_i, а $X_{ост\_t}(i)$ – вектор размерности $p - p_i$ предопределенных переменных, не входящих в уравнение.

Критерий переопределения позволяет проверить гипотезу о том, что рассматриваемое структурное уравнение (3) не содержит предопределенных переменных $X_{ост\_t}(i)$, на основе статистики [2]:

$$n\left[\frac{r}{\bar{\gamma}^T(i)M_{EE}\bar{\gamma}(i)} - 1\right], \quad (4)$$

где $r = \bar{\gamma}^T(i)M_{Y_iY_i}\bar{\gamma}(i) - \bar{h}\_^T(i)M_{X(i)X(i)}\bar{h}\_(i)$, E – случайные ошибки приведенной формы, $W_t = X_t(i) - M_{X_{ост}(i)X(i)}M^{-1}_{X(i)X(i)}X_t(i)$, $\bar{\gamma}(i)$ и $\bar{h}\_(i)$ – оценки параметров, $M_{zz} = \frac{1}{n}\sum_{t=1}^{n} z_t z_t^T$, а

$$M_{EE} = M_{Y_iY_i} - M_{Y_iX(i)}M^{-1}_{X(i)X(i)}M_{X(i)Y_i} - \\ - M_{Y_iW}M^{-1}_{WW}M_{WY_i}.$$

Для нахождения оценок параметров $\bar{\gamma}(i)$ и $\bar{h}\_(i)$ можно использовать любой из состоятельных методов, таких как двухшаговый МНК (TSLS) [3], метод Комиссии Коулса (LIML) [2], модифицированный LIML SDUNB [5].

При верной гипотезе статистика (4) асимптотически подчиняется χ^2-распределению со степенью свободы $p - p_i - m_i + 1$.

Как уже было сказано, распределение статистики (2) в [1] рассматривается только при нормальном распределении шума, а в [2] нет информации о том, при каком распределении шума статистика (4) подчиняется своему асимптотическому распределению.

Точечный прогноз эндогенных переменных $\bar{Y}_{n+\tau}$ на τ временных тактов вперед строится с помощью формулы:

$$\bar{Y}_{n+\tau} = \bar{\varPi}X_{n+\tau}, \quad (5)$$

где $X_{n+\tau}$ – значения предопределенных переменных в момент времени $n+\tau$, $\bar{\varPi}$ – оценка матрицы приведенной формы СОУ, которую можно получить либо по обычному методу наименьших квадратов (OLS), либо для ее нахождения использовать оценки матриц структурной формы, полученные по состоятельным методам (TSLS, LIML, SDUNB).

Истинные значения эндогенных переменных $Y_{n+\tau}$ с заданной доверительной вероятностью $1-\alpha$ концентрируются внутри эллипсоида рассеяния с центром в точке $\bar{Y}_{n+\tau}$, вывод которого опирается на тот факт, что при нормально распределенных случайных составляющих статистика Хотеллинга подчиняется распределению Фишера со степенями свободы m и $n-p-m+1$ при условии, что среди предопределенных переменных нет лаговых эндогенных.

Статистика Хотеллинга имеет следующий вид [4]:

$$\frac{n-p-m+1}{(n-p)m}T, \qquad (6)$$

где $T=(\bar{Y}_{n+\tau}-Y_{n+\tau})^T\,\bar{\Sigma}^{-1}_{\bar{\varepsilon}(\tau)}(\bar{Y}_{n+\tau}-Y_{n+\tau})$, а $\bar{\Sigma}^{-1}_{\bar{\varepsilon}(\tau)}$ – оценка ковариационной матрицы ошибок прогноза $\bar{\varepsilon}(\tau)=\bar{Y}_{n+\tau}-Y_{n+\tau}$.

IV. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Объем N выборок значений исследуемых величин (статистик, ошибок прогноза, оценок) принимался равным 10000.

Рассмотренные системы:

$$\begin{cases} y_t^{(1)}=-\gamma_{12}y_t^{(2)}+h_{11}x_t^{(1)}+\delta_t^{(1)}, \\ y_t^{(2)}=-\gamma_{21}y_t^{(1)}+h_{22}x_t^{(2)}+\delta_t^{(2)}, \end{cases} \qquad (7)$$

$$\begin{cases} y_t^{(1)}=-\gamma_{12}y_t^{(2)}+h_{11}x_t^{(1)}+h_{12}x_t^{(2)}+\delta_t^{(1)}, \\ y_t^{(2)}=-\gamma_{21}y_t^{(1)}+h_{23}x_t^{(3)}+h_{24}x_t^{(4)}+\delta_t^{(2)}, \end{cases} \qquad (8)$$

Предполагалось, что истинные значения параметров системы (7) равны $\gamma_{12}=2$, $h_{11}=3$, $\gamma_{21}=-2$, $h_{22}=4$, а системы (8) – $\gamma_{12}=1$, $h_{11}=3$, $h_{12}=2$, $\gamma_{21}=-2$, $h_{23}=4$, $h_{24}=1$. Предопределенные (экзогенные) переменные изменялись в диапазоне [-5,5].

Исследования распределений статистик (2), (4), (6) проводились на системах (7), (8) для ряда объемов выборок n (малого размера и достаточно большого) и случайных составляющих, распределенных по экспоненциальному семейству с плотностью

$$\frac{\lambda}{2\theta_1\Gamma(1/\lambda)}\exp\left\{-\left(\frac{|x-\theta_0|}{\theta_1}\right)^{\lambda}\right\}, \qquad (9)$$

и значениями параметра формы $\lambda=2$ (нормальное распределение), $\lambda=1$ (распределение Лапласа), $\lambda=4$, $\lambda=0.5$ и $\lambda=0.25$.

Из результатов моделирования следует, что при нормальном шуме распределение статистики Андерсона-Рубина достаточно хорошо согласуется с распределением $F(p-p_i,n-p)$, как при малых объемах выборок, так и при больших.

Кроме того, моделирование показало, что отклонения распределения шума от нормального закона в сторону Лапласа и семейства (9) с $\lambda=4$ не оказывает сильного влияния на распределение статистики (2).

Но отклонения распределения шума от нормального закона в сторону семейства (9) с $\lambda=0.5$ и $\lambda=0.25$ оказывают более заметное влияние на распределение статистики Андерсона-Рубина при малых объемах выборок.

Так, на рис. 1 представлено эмпирическое и теоретическое распределение статистики Андерсона-Рубина для первого уравнения системы (7) при объеме выборки 15 и случайных составляющих, подчиненных закону Лапласа. Графики визуально практически совпадают, а количественной мерой близости служат достигнутые уровни значимости по критериям согласия, которые в данном случае достаточно высокие.

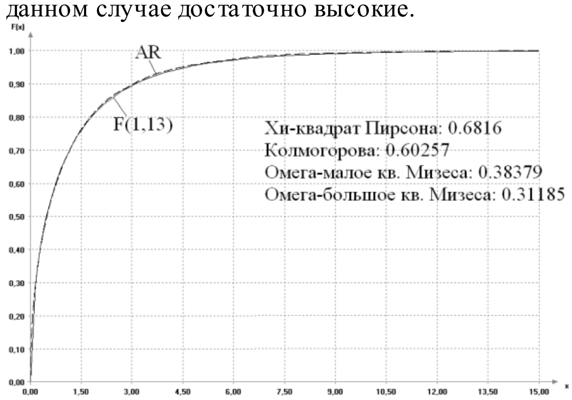

Рис. 1. Распределение статистики Андерсона-Рубина при n=15 и шуме, подчиненному закону Лапласа

На рис. 2 представлено эмпирическое распределение статистики Андерсона-Рубина для первого уравнения системы (7) при случайных составляющих, имеющих распределение (9) с $\lambda=0.25$, и объеме выборки 15 и соответствующее распределение Фишера. Графики заметно отличаются визуально, а достигнутые уровни значимости по всем критериям согласия равны нулю.

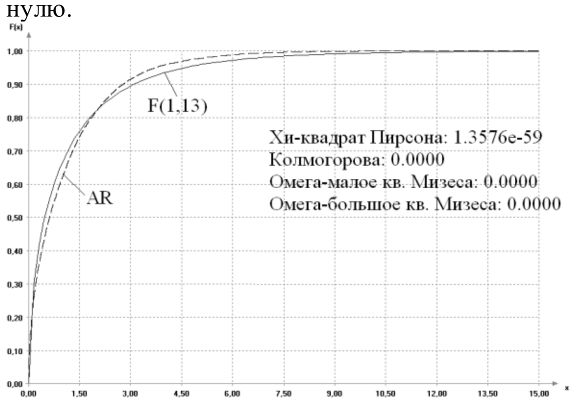

Рис. 2. Распределение статистики Андерсона-Рубина при шуме, распределенном по экспоненциальному семейству с параметром формы 0.25, и n=15

Критерий переопределения показал большую устойчивость к отклонениям распределения шума от нормального закона. Графики эмпирического распределения статистики (4) в этом случае практически сливаются, поэтому приводить их не имеет смысла.

Но, в отличие от статистики Андерсона-Рубина, распределение статистики критерия переопределения при малом объеме выборки ощутимо отличается от асимптотического распределения при нормальном законе. Так, на рис. 3 представлено теоретическое и эмпирическое распределение статистики (4) для первого уравнения системы (8) при разных объемах выборки. В данном случае при объеме выборки 70 гипотеза о согласии с теоретическим законом распределения не отвергается.

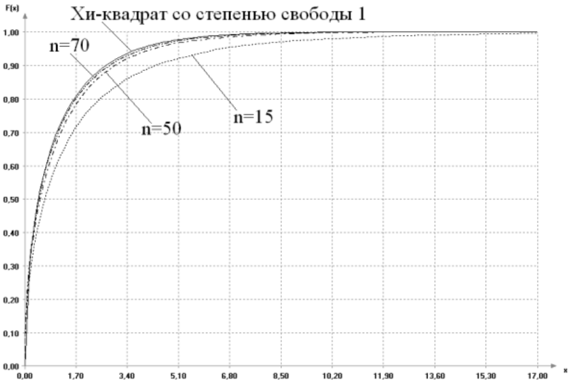

Рис. 3. Распределение статистики критерия переопределения при разном объеме выборки

Метод оценивания параметров практически не влияет на распределение статистики критерия переопределения.

Когда в точечном прогнозе (5) используется оценка матрицы приведенной формы по методу OLS (Reduce OLS) и среди предопределенных переменных нет лаговых эндогенных, распределение статистики Хотеллинга при нормальном шуме хорошо согласуется с $F(m, n-p-m+1)$ как при малых объемах выборок, так и при больших.

Но распределение статистики Хотеллинга очень чувствительно к отклонениям распределения шума от нормального закона. На рис. 4 приведено эмпирическое распределение статистики (6) для системы (8) при объеме выборки 15 и случайных составляющих, распределенных по нормальному закону, экспоненциальному семейству (9) с $\lambda = 4$ и закону Лапласа.

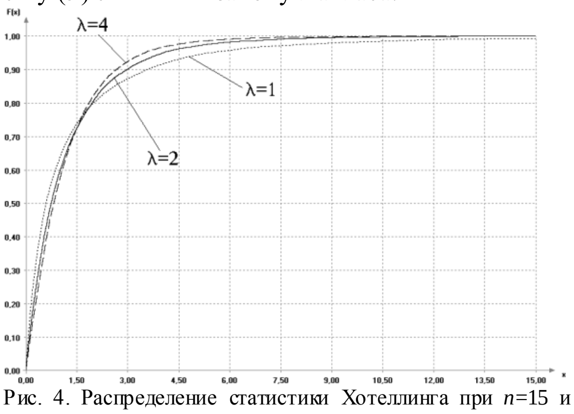

Рис. 4. Распределение статистики Хотеллинга при n=15 и разном распределении шума

Из рис. 4 видно, что использование верхних 100α-процентных точек «классического» распределения Фишера в случае экспоненциального семейства (9) с $\lambda = 4$ будет приводить к интервалам, покрывающим истинные значения эндогенных переменных с вероятностью большей, чем $1-\alpha$, а в случае распределения Лапласа – к интервалам, покрывающим истинные значения эндогенных переменных с вероятностью меньшей, чем $P = 1-\alpha$. Как показало моделирование, для случая, представленного на рис. 4, при использовании 10% точки распределения $F(2,10)$ (распределения Фишера, соответствующего данному случаю) вероятность покрытия истинных значений эндогенных переменных для нормального закона равна 0.8951, для экспоненциального семейства (9) с $\lambda = 4$ – 0.9186, а для закона Лапласа – 0.868.

Чтобы при нарушении предположения нормальности шума правильно построить интервальный прогноз можно для нахождения верхней процентной точки использовать либо непосредственно полученную в процессе моделирования эмпирическую функцию распределения статистики (6), либо приближенную аналитическую модель, хорошо ее аппроксимирующую. Исследования показали, что распределение статистики Хотеллинга при отклонениях распределения шума от нормального закона наилучшим образом описывается бета-распределением 2-го или 3-го рода и Г-распределением.

Из результатов исследования следует, что если среди предопределенных переменных присутствуют лаговые эндогенные переменные, принимающие действительно наблюденные значения, то ситуация при построении интервального прогноза ничем не отличается от рассмотренных выше. Если же лаговые эндогенные переменные, входящие в СОУ, также приходится прогнозировать, то распределение статистики Хотеллинга отличается от ее распределений в рассмотренных ранее случаях.

Если для нахождения оценки матрицы приведенной формы сверхидентифицируемой системы использовать оценки матриц структурной формы, полученные по состоятельным методам (TSLS, LIML, SDUNB), то эта оценка при малых объемах выборок будет точнее, чем при использовании метода Reduce OLS. Это связано с тем, что метод Reduce OLS не учитывает сверхидентифицирующие ограничения, наложенные на приведенную форму системы. Следовательно, в данном случае и точечные прогнозы, построенные при использовании оценки матрицы приведенной формы, полученной из оценок матриц структурной формы по методам TSLS, LIML, SDUNB (Reduce SDUNB), при малых объемах выборок будут точнее, чем при использовании метода Reduce OLS.

Так, на рис. 5 представлены эмпирические распределения ошибки прогноза эндогенной переменной $y_t^{(2)}$ системы (8) при оценках Reduce OLS и Reduce SDUNB, объеме выборки 15 и нормально распределенных случайных составляющих.

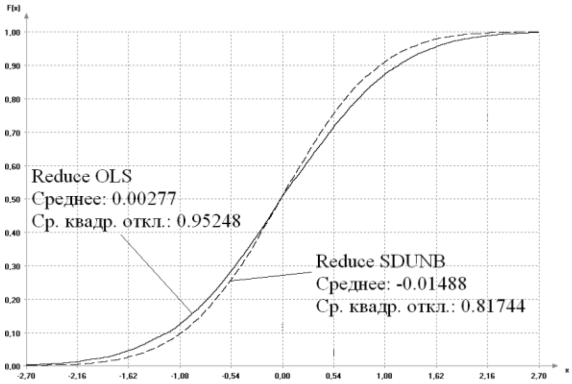

Рис. 5. Распределение ошибки прогноза при n=15 и разных методах оценивания приведенной формы

Из рис. 5 видно, что средние значения ошибок прогноза близки к нулю, что говорит о несмещенности прогнозов, а среднее квадратическое отклонение для оценки Reduce SDUNB меньше, чем для оценки Reduce OLS. Таким образом, при малых объемах выборок для сверхидентифицируемой системы прогноз, основанный на оценках Reduce OLS, менее точен, чем прогноз, основанный на оценках Reduce SDUNB.

С ростом объема выборки различие между оценками матрицы приведенной формы становится менее существенным, а, следовательно, и различие распределений ошибок прогноза также становится менее существенным.

Использование верхних процентных точек действительного распределения статистики Хотеллинга, вычисленной при оценке Reduce SDUNB, для построения интервальных прогнозов для сверхидентифицируемых систем при малых объемах выборок будет приводить к более узким интервалам, чем при оценке Reduce OLS, при той же вероятности $1-\alpha$.

V. ЗАКЛЮЧЕНИЕ

Методика компьютерного моделирования позволяет относительно просто и быстро исследовать статистические закономерности. Конечно, она имеет и свои недостатки. Выводы, полученные по методике компьютерного моделирования, имеют меньшую степень обобщения по сравнению с аналитическими.

В работе методика компьютерного моделирования была успешно применена к исследованию распределений статистик критериев и процедуры построения прогнозов эндогенных переменных.

По результатам исследования можно сделать следующие выводы.

Если распределение шума имеет не слишком «тяжелые» хвосты по сравнению с нормальных законом, то критерий Андерсона-Рубина можно без особых опасений применять на практике, как при малых объемах выборок, так и при больших.

Асимптотическое распределение статистики критерий переопределения можно применять без опасения сделать ошибочный вывод только при достаточно больших объемах выборки.

При нормальном распределении шума построение интервальных прогнозов эндогенных переменных при использовании оценок Reduce OLS будет корректным при любом объеме выборки.

В случае отклонения распределения шума от нормального закона можно построить приближенную аналитическую модель, хорошо описывающую распределение статистики Хотеллинга, и пользоваться ей для нахождения верхних процентных точек.

Лаговые эндогенные переменные не влияют на распределение статистики Хотеллинга, если они принимают действительно наблюденные значения.

Использование оценок Reduce SDUNB и др. для сверхидентифицируемых систем обеспечивает при малых объемах выборок более точные прогнозы, чем использование оценок Reduce OLS.

СПИСОК ЛИТЕРАТУРЫ

[1] Dufour Jean-Marie. Identification, weak instruments and statistical inference in econometrics. Université de Montréal, 2003.

[2] Маленво Э. Статистические методы в эконометрии. / Пер. с франц. - М.: Статистика, 1976, вып. 2.

[3] Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. Учебник для вузов.- М.: ЮНИТИ, 1998. - 1022 с.

[4] Hooper J.W., Zellner A. The error of forecast for multivariate regression model / Econometrica, 1961. Vol. 29, pp. 544-555.

[5] Dennis Oberhelman, K. Rao Kadiyala. An empirical investigation of the distribution of modified LIML estimators for equations with more than two endogenous variables. Sankhya: The Indian Journal of Statistics, 1999. Vol. 61, Series B, Pt. 2, pp. 518-532.

Лемешко Борис Юрьевич	Профессор кафедры прикладной математики Новосибирского государственного технического университета (НГТУ), д.т.н., профессор
Щеглов Алексей Евгеньевич	Магистр прикладной математики и информатики, в 2008 г. закончил НГТУ.