MATHCAD В РУКАХ ЭКОНОМИСТА: БОКС-КОКС ПРЕОБРАЗОВАНИЕ И ИЛЛЮЗИЯ «НОРМАЛЬНОСТИ» МАКРОЭКОНОМИЧЕСКОГО РЯДА

А.Н. Порунов,
кандидат экономических наук, научный сотрудник лаборатории стратегических исследований и операционного проектирования Самарского государственного технического университета,
e-mail: rameno@rambler.ru.
Адрес: г. Рамено, Сызранский район, Самарская область, ул. Пионерская, д. 5.

В статье рассматривается методика преобразования в среде Mathcad ненормально распределенного ряда макроэкономического ряда к нормально распределенному на основе преобразования Бокса-Кокса и возникающие при этом ошибки в оценке нормальности распределения.

Ключевые слова: преобразование Бокса-Кокса, макроэкономический ряд, непараметрические методы, параметрические методы, робустные методы.

Введение

Очень часто[1] экономисту-аналитику приходиться иметь дело со статистическими данными, которые по тем или иным причинам не проходят тест на нормальность. В этой ситуации есть два выхода: либо обратиться к непараметрическим методам, что весьма проблематично для экономиста, поскольку требует изрядной математической подготовки, либо воспользоваться специальными методами, позволяющими преобразовать исходную «ненормальную статистику» в «нормальную», что само по себе так же непросто.

Широко распространено мнение, что если же данных много (например, $n > 100$), или исследуются переменные, значения которых определяются бесконечным числом независимых факторов, то не имеет смысла использовать непараметрические статистики и в этой ситуации лучше обратиться к методам трансформации ненормально распределенных данных в нормально распределенные. Среди множества таких методов преобразований одним из лучших (при неизвестном типе распределения) считается Бокс-Кокс преобразование.

Авторы этого преобразования известные статистики – Джордж Эдвард Пелхэм Бокс (George

[1] Математики-экономисты считают, что «очень часто» мягко сказано, здесь следовало бы сказать «в абсолютном большинстве случаев».

Edward Pelham Box), профессор Висконсинского университета в городе Мэдисон (США) и сэр Дэвид Роксби Кокс (Sir David Roxbee Cox) – профессор колледжа Бирбека лондонского университета. Впервые, суть предлагаемого метода была изложена ими в 1964 году, в Журнале Королевского статистического общества (GB) [1]. Практические аспекты Бокс-Кокс преобразования (БК), сегодня достаточно подробно рассмотрены в специальной англоязычной литературе [2–7], чего нельзя сказать об отечественной. Рассмотрим, так ли всемогуще БК преобразование в борьбе с «ненормально» распределенным макроэкономическим рядом и какие иллюзии могут возникнуть у исследователя-экономиста, в зависимости от степени его «статистической испорченности» при оценке согласия функций эмпирического и теоретического распределений.

Бокс-Кокс преобразование

Пусть некоторая, непрерывная во времени, функция X представлена вектором её значений x_i, $i \in 1, \ldots N$. Бокс-Кокс преобразование определяется следующим образом:

$$x(\lambda) = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda}, \lambda \neq 0 \\ \ln(x), \lambda = 0 \end{cases} \quad (1)$$

Выражение (1) представляет собой универсальное параметрическое семейство преобразований, которое экономисты часто используют в алгоритмах сезонной (циклической) корректировки, для того чтобы сезонная составляющая преобразованного динамического ряда стала (хотя бы в первом приближении) не эволюционирующей по амплитуде, что упрощает ее последующую идентификацию [3]. Тиражируемые в литературе по экономической статистике и по этой причине популярные среди экономистов, логарифмическое и степенное преобразования, представляют лишь частный случай преобразования БК. Так, например, в зависимости от значений λ получаем: при λ=0 – логарифмическое, при λ<>2 – степенное преобразование.

Один из способов выбрать оптимальное значение λ, — это использование значения λ, максимизирующего логарифм функции правдоподобия.

Логарифм функции правдоподобия:

$$f(x,\lambda) = -\frac{N}{2} \cdot \ln\left[\sum_{i=1}^{N} \frac{(x_i(\lambda) - \bar{x}(\lambda))^2}{N}\right] + (\lambda - 1) \cdot \sum_{i=1}^{N} \ln(x_i) \quad (2)$$

где $x(\lambda) = \dfrac{1}{N} \cdot \sum_{i=1}^{N} x_i(\lambda)$ –

есть среднеарифметическая БК преобразованных данных.

Поскольку изначально БК преобразование было ориентировано только на положительные величины, проблему учета отрицательных значений данных снимают, добавляя к исходным значениям некоторое смещение, переводящее все отрицательные величины в положительную область[2]:

$$x(\lambda) = \begin{cases} \dfrac{(x+c)^{\lambda} - 1}{\lambda}, \lambda \neq 0 \\ \ln(x+c), \quad \lambda = 0 \end{cases} \quad (3)$$

где: c – величина смещения.

При этом должно выполняться условие:

$$(x+c) > 0 \quad \forall x_i \in X$$

Доверительная оценка λ (с использованием статистики отношения правдоподобия) может быть произведена следующим образом:

$$f(x,\lambda) \geq f(x,\tilde{\lambda}) - 0.5\chi^2_{\alpha,1}, \quad (4)$$

где $\tilde{\lambda}$ – оценка максимального правдоподобия для λ;

$\chi^2_{\alpha,1}$ – верхняя 100(1–α) процентиль хи-квадрат распределения с 1-ой степенью свободы.

Практическая реализация

Для иллюстрации процедуры БК преобразования в среде Mathcad[3] использовался таблично заданный, макроэкономический ряд ВВП РФ – ряд X *(табл.1)*.

[2] таким образом получается двухпараматрическое семейство преобразований которое сегодня называется преобразованием Бокса-Кокса

[3] В большинстве современных математических пакетов сдвиг на константу (смещение) не предусмотрен, т.е. используется алгоритм более простого однопараметрического преобразования.

Таблица 1.

Динамика уровней ВВП РФ за период 1885–2009 гг.[4]

T	xt	t	xt	t	xt	t	xt
1885	76	1917	143	1949	301	1981	1440
1886	73	1918	116	1950	374	1982	1423
1887	80	1919	92	1951	440	1983	1477
1888	86	1920	77	1952	453	1984	1664
1889	79	1921	74	1953	476	1985	1661
1890	75	1922	69	1954	483	1986	1668
1891	65	1923	64	1955	536	1987	1666
1892	93	1924	82	1956	569	1988	1754
1893	92	1925	98	1957	610	1989	1763
1894	95	1926	121	1958	616	1990	1784
1895	106	1927	146	1959	692	1991	1745
1896	93	1928	162	1960	721	1992	1735
1897	105	1929	173	1961	691	1993	1716
1898	94	1930	152	1962	789	1994	1518
1899	89	1931	175	1963	830	1995	1282
1900	90	1932	166	1964	818	1996	1267
1901	87	1933	171	1965	849	1997	1141
1902	86	1934	208	1966	958	1998	1119
1903	99	1935	242	1967	970	1999	1156
1904	95	1936	293	1968	1020	2000	1068
1905	114	1937	289	1969	1062	2001	1111
1906	98	1938	295	1970	1047	2002	1173
1907	88	1939	333	1971	1086	2003	1332
1908	89	1940	359	1972	1203	2004	1585
1909	108	1941	382	1973	1273	2005	1746
1910	111	1942	344	1974	1218	2006	1860
1911	123	1943	225	1975	1253	2007	2001
1912	107	1944	202	1976	1349	2008	2271
1913	118	1945	217	1977	1420	2009	2074
1914	134	1946	194	1978	1469		
1915	158	1947	225	1979	1466		
1916	160	1948	280	1980	1424		

Для нахождения уравнения тренда (в случае экспоненциальной зависимости) воспользуемся стандартной, встроенной в Mathcad[5] функцией *exp fit (t, X, g)*. Эта функция возвращает вектор, содержащий три коэффициента экспоненциальной кривой вида: $a \cdot exp(b \cdot x) + c$, которая наилучшим образом аппроксимирует данные в векторах *t* и *X*. Необязательный вектор *g* содержит начальное приближение для этих трех коэффициентов:

$$g = \begin{pmatrix} 0.001 \\ 0.001 \\ 0.001 \end{pmatrix}$$

$$c = exp\ fit\ (t, X, g)$$

$$c = \begin{pmatrix} 0.00000000075 \\ 0.0143518673 \\ -508.15140440551 \end{pmatrix}$$

$$Xtrend = (c_1 \cdot exp\ (c_2 \cdot t) + c_3$$

Для приведения ряда к стационарному виду из ряда *X* вычитают найденный тренд – *Xtrend* и определяют ряд остатков ΔR *(рис.2)*:

$$\Delta R = X - Xtrend$$

Для проверки близости распределения ряда остатков к нормальному распределению, построим гистограмму распределения *H (рис.3)*, используя функцию

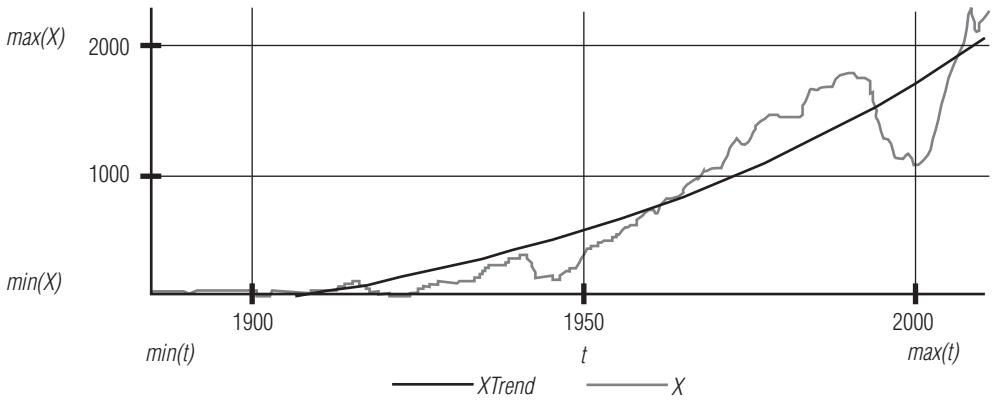

Рис.1. Динамический ряд X и тренд

[4] в современных границах РФ, составлен автором по источникам [8-12]

[5] Использовалась последняя модифицированная версия пакета Mathcad-14 M-035

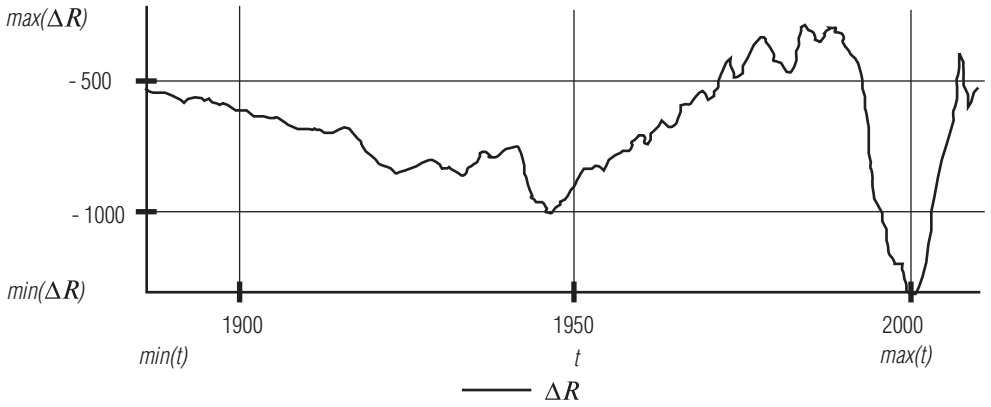

Рис.2. Динамика ряда остатков ΔR

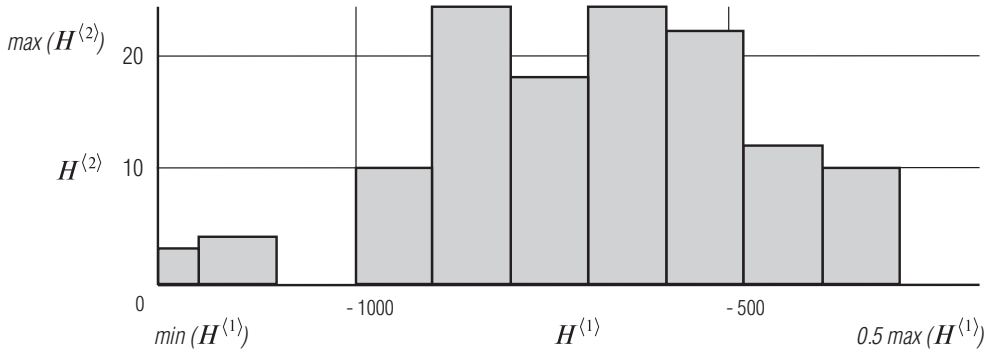

Рис.3. Гистограмма распределения ряда остатков

$$H = histogram\ (trunc\ (\sqrt{N}, \Delta R),$$

где: *histogram (trunc* $(\sqrt{N}, \Delta R)$ – функция, возвращающая матрицу H из двух столбцов, содержащую средние точки *trunc* $\sqrt{N}$ подинтервалов. Результирующая матрица содержит *trunc* $\sqrt{N}$ строк, где *trunc* – функция, возвращающая целую часть аргумента.

Как видно из гистограммы, характер распределения ряда остатков далёк от нормального. Как показывает практика, может оказаться, «...что преобразование квадратного корня еще слабовато (не поджимает справа хвост распределения), а логарифмическое – уже слишком сильное (хвостик появляется слева). Раньше пришлось бы выбирать из этих двух, но преобразование Бокса-Кокса в этом случае (λ между 0 и 0,5) найдет промежуточное решение. Поэтому, если истинное нормализующее преобразование неизвестно, преобразование Бокса-Кокса считается лучшим» [13].

Поскольку БК преобразование применяется только к положительным уровням ряда, выберем величину смещения так, чтобы $(\Delta R + c) > 0$ при любых значениях ряда остатков ΔR. Примем величину смещения несколько большей (для наглядности, – на 20%) минимального значения в ряду остатков ΔR: $c = 1.2min\ (\Delta R)$.

Тогда новый ряд остатков ΔRg, с учетом смещения, будет равен:

$$\Delta Rg = \Delta R - 1.2 \cdot min(\Delta R)$$

где: $min(\Delta R)$ – функция, возвращающая наименьшее из значений ΔR.

Пусть показатель степени изменяется в пределах: $\lambda = -1, -1 + 0.1 \ldots 15$ с шагом 0.1, тогда лог-функцию правдоподобия $FP(\Delta Rg, \lambda)$ можно определить следующим образом:

$$FP(\Delta Rg, \lambda) = \frac{-N}{2} \cdot ln\left[\sum_{i=1}^{N} \frac{\left[\frac{(\Delta Rg_i)^{\lambda} - 1}{\lambda} - \frac{1}{N} \cdot \sum_{i=1}^{N} \frac{(\Delta Rg_i)^{\lambda} - 1}{\lambda}\right]^2}{N}\right] +$$

$$+ (\lambda - 1) \cdot \sum_{i=1}^{N} ln(\Delta Rg_i)$$

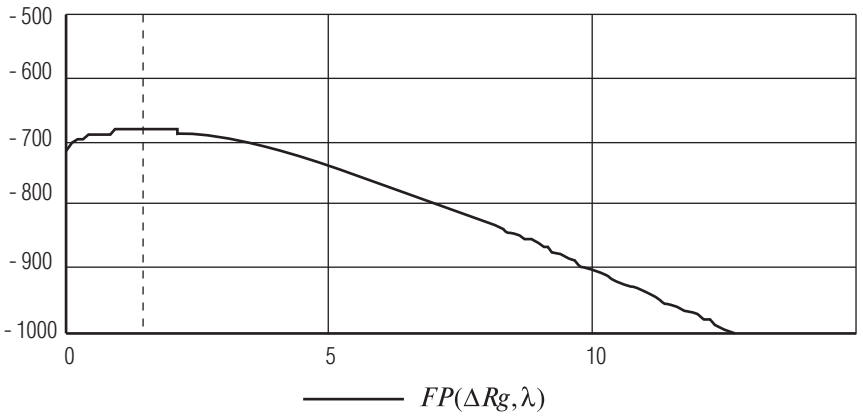

Рис.4. График логарифмической функции правдоподобия

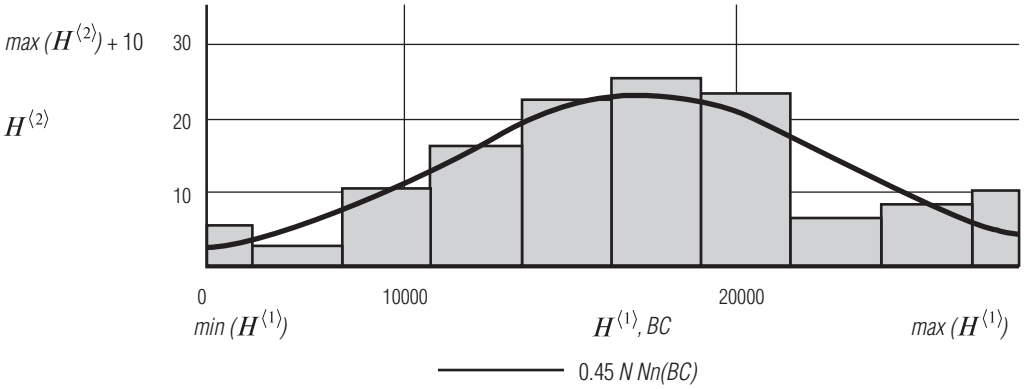

Рис. 5. Гистограмма ряда остатков после ВК преобразования

Для того чтобы найти оптимальное значение λ_{opt}, итеративно подставляем значения λ при которых логарифмическая функция правдоподобия $FP(\Delta Rg, \lambda)$ достигает максимума. Ориентируясь по графику логарифмической функции правдоподобия, возьмем «вилку» из значений:

$$FP(\Delta Rg, 1.48) = -682.903$$
$$FP(\Delta Rg, 1.49) = -682.902$$
$$FP(\Delta Rg, 1.50) = -682.903$$

Промежуточное значение $FP(\Delta Rg, 1.49)$ соответствует максимуму функции $FP(\Delta Rg, \lambda)$ т.е. в данном случае $\lambda_{opt} = 1.49$

Тогда преобразованный ряд остатков *BC*, будет определяться по формуле:

$$BC = \frac{\Delta Rg^{1.49} - 1}{1.49}$$

Определим еще один ряд ΔRn, получаемый в результате сортировки ряда остатков *BC*:

$$\Delta Rn = sort(BC),$$

где: *sort(BC)* – функция, возвращающая вектор со значениями из *BC*, упорядоченными по возрастанию.

Это позволит нам отразить кривую плотности нормального распределения на гистограмме *(рис.5)*:

$$H = histogram(trunc(\sqrt{N} - 1, BC)$$

Классическая форма функции плотности нормального распределения (гаусиан) в принятых обозначениях будет иметь следующий вид:

$$Nn(\Delta Rn) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot exp\left[\frac{-1}{2} \cdot \left(\frac{\Delta Rn - mean(\Delta Rn)}{Stdev(\Delta Rn)}\right)^2\right],$$

где: $mean(\Delta Rn)$ – функция, возвращающая арифметическое среднее (среднее значение) элементов ΔRn; $Stdev(\Delta Rn)$ – функция, возвращающая среднеквадратическое отклонение совокупности элементов ΔRn.

Гистограмма *(рис.5)* показывает, что характер распределения остатков, после преобразования по методу Бокса-Кокса, близок к нормальному. «За-

быв» о критериях согласия, оценим ряд остатков на нормальность распределения, на основе показателей эксцесса и асимметрии. Коэффициент асимметрии: $skew(BC) = -0.0334$, где: $skew(BC)$ –функция, возвращающая асимметрию элементов BC. Эксцесса: $kurt(BC) = -0.01163$, где: $kurt(BC)$ функция, возвращающая асимметрию элементов BC.

Рассчитаем вспомогательные величины σA и σE:

$$\sigma A = \sqrt{\frac{6 \cdot (N-2)}{(N+1) \cdot (N+3)}} = 0.2123$$

$$\sigma E = \sqrt{\frac{24 \cdot N \cdot (N-2) \cdot (N-3)}{(N+1)^2 \cdot (N+3) \cdot (N+5)}} = 0.4099$$

Для ряда с распределением близким к нормальному должны выполняться следующие условия [12]:

$$|skew(BC)| = 0.0334 < 1.5 \cdot \sigma \cdot A = 0.3185$$

и $$\left| kurt(BC) - \frac{6}{N+1} \right| = 0.16 < 1.5 \cdot \sigma \cdot E = 0.6149 .$$

В данном случае эти условия выполняются. Продолжим проверку. С этой целью проведем, очень популярный сегодня у экономистов, визуальный анализ нормальности. Стандартизируем, сортированный ранее ряд остатков ΔR, предполагая, что справедлива гипотеза о нормальности ряда:

$$BSn = \frac{\Delta R - mean(\Delta R)}{Stdev(\Delta R)}$$

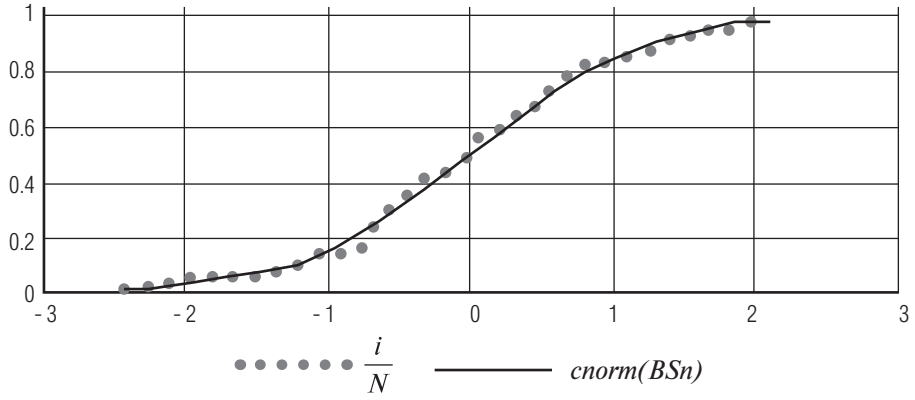

Рис. 6. Графики эмпирической и теоретической функций распределения

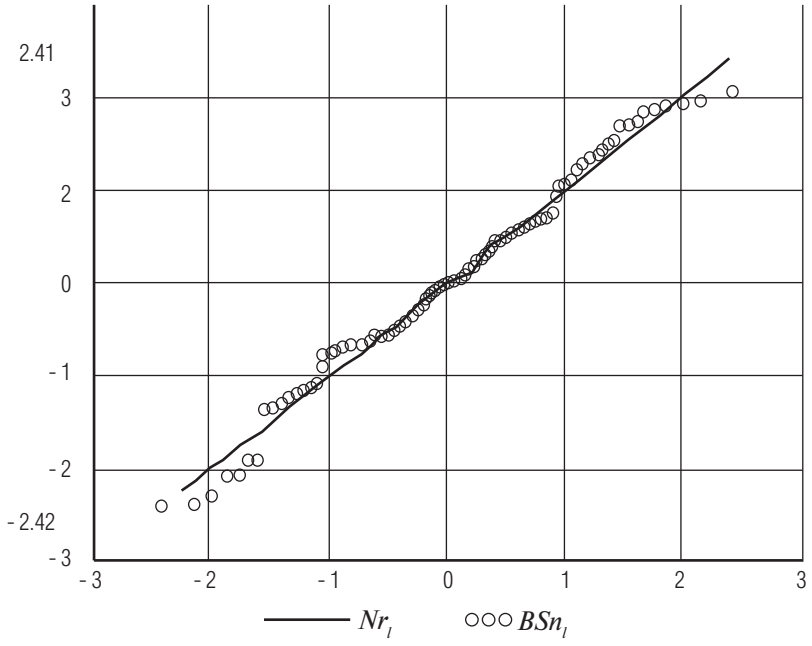

Рис. 7. Графики обратных кумулятивных распределений эмпирической и теоретической функций

Построим эмпирическую функцию распределения $\frac{i}{N}$ и сравним её с теоретическим распределением *(рис. 6),* используя встроенную mathcad-функцию *cnorm(BSn)*. Эта функция возвращает кумулятивное распределение вероятностей со средним, равным 0 и дисперсией, равной 1:

График *(рис. 6)* показывает близость кривых распределения $\frac{i}{N}$ и *cnorm(BSn)*. На основе mathcad-функции, *gnorm(F,* μ, σ*)* возвращающую обратное кумулятивное нормальное распределение ряда *F* с заданными средним μ и среднеквадратическим отклонением σ, построим еще один график зависимости $BSn(Nr_l)_l$ *(рис. 7).* Предварительно определим:

$$1=1...N\text{-}1,\ F_l = \frac{l}{N},\ Nr_l = qnorm\,(F_l,\ 0,1)$$

На первый взгляд может показаться, что и р*ис. 7* не дает оснований для беспокойства, – большая часть точек стандартизированного ряда остатков *BSn* располагаются очень близко к прямой, и, поэтому, распределение ряда можно считать нормальным. Подобные заключения не редки в работах, посвященных исследованию макроэкономических рядов. Но самое печально то, что множатся случаи, когда этим и ограничивается процедура проверки гипотезы о нормальности распределения. Тем временем использование уже старого, «доброго» критерия согласия Пирсона (в данном случае, при *N*=127 его использование оправдано), критерия Колмогорова или омега-квадрат говорит, что «не все спокойно в датском королевстве». Покажем, так ли это? Тем более, что Mathcad позволяет это сделать достаточно просто (для понимания) и наглядно.

Для начала рассчитаем критерий Пирсона. С этой целью определим размах вариации стандартизированного ряда остатков:

$$R = BSn_N - BSn_1 = 4.5\,.$$

Проведем группировку ряда, число групп:

$$K = trunc(\sqrt{N}\text{-}1)=10,\quad k=1...K\,.$$

Величина интервала :

$$h = \frac{R}{K} = 0.45.$$

Середины интервалов:

$$m_k = BSn_k + \frac{h}{2} + (k\text{-}1)\cdot h\,,$$

m_k=(-2.2 -1.71 -1.18 -0.51 -0.04 0.56 1.02 2 2.48 2.97)

Рассчитаем теоретические частоты f_k:

$$f_k = h\cdot N\cdot \frac{1}{\sqrt{2\cdot\pi}}\cdot e^{\frac{-1}{2}(m_k)^2}\,,$$

$$f_k = (2\quad 5\quad 11\quad 20\quad 23\quad 20\quad 14\quad 3\quad 1\quad 0),$$

и эмпирические частоты (используем определенные ранее данные для построения гистограммы *(рис. 5)*:

$$H^{(2)} = (5\ \ 2\ \ 10\ \ 16\ \ 25\ \ 23\ \ 6\ \ 8\ \ 10),$$

тогда расчетный критерий Пирсона χ^2 будет равен:

$$\chi^2 = \sum_k \frac{\left[f_k - \left(H^{\langle 2\rangle}\right)_k\right]^2}{\left(H^{\langle 2\rangle}\right)_k} = 30.3.$$

При уровне значимости α=0.05 и числе степеней свободы $s = K - 3 = 7$, табличное значение критической точки правосторонней критической области $\chi^2_{кр} = 14.2$. Таким образом, эмпирические и теоретические частоты отличаются значимо.

Далее определим значения статистики Колмогорова:

$$KBSn_i = \left|\frac{i}{N} - cnorm(BSn_i)\right|$$

где: $cnorm(BSn_i)$ – mathcad функция возвращающая кумулятивное распределение вероятностей со средним, равным 0, и дисперсией, равной 1.

Статистика Колмогорова

$$D = max(KBSn) = 0.06\,.$$

Расчетное значение статистики:

$$Kt = \sqrt{N}\cdot D = 0.71,$$

при выбранном уровне значимости α=0.05 превышает табличное значение

$$\frac{1.36}{\sqrt{N}} = 0.12\,,$$

это означает, что нулевую гипотезу следует отвергнуть, т.е. характер распределения ряда остатков далек от нормального, несмотря на проведенное ранее его БК преобразование.

Заключение

Практика статистических исследований показывает, «...что распределения реальных данных никогда не входят в какое-либо параметрическое семейство» [14]. Сегодня в статистической литературе есть немало примеров, показывающих, что распределения ошибок измерений почти всегда отличаются от нормальных» [15]. Эти семейства – лишь возможные приближения, которые далеко не всегда являются адекватными. Приведенный выше анализ конкретных данных приводит к аналогичному заключению.

В этой связи нельзя не согласиться с мнением одного из авторитетных отечественных статистиков – профессора А.И. Орлова, о том, что не умаляя значимости методов параметрической статистики, необходимо переходить к непараметрическим и робастным методам [14]. И, в первую очередь, по мнению автора, это относится к исследованию макроэкономических рядов. Экономистам об этом надо помнить. ■

Литература

1. Box, G. E. P.; Cox, D. R. An analysis of transformations. (With discussion) J. Roy. Statist. Soc. Ser. B 26 1964 211–252. http://www.ams.org/mathscinet-getitem?mr=192611
2. Box-Cox Transformations: An Overview. Pengfei Li. Department of Statistics, University of Connecticut. Apr 11, 2005 http://www.stat.uconn.edu/~studentjournal/index_files/pengfi_s05.pdf
3. Carroll, RJ and Ruppert, D. On prediction and the power transformation family. Biometrika 68: 609–615.
4. Box-Cox Transformation. http://www-stat.stanford.du/~olshen/manuscripts/selenite/node6.html
5. Davidson, Russell, and James G. MacKinnon. 1993. Estimation and Inference in Econometrics. Oxford University Press.
6. Definition of Box-Cox Transformation http://economics.about.com/cs/economicsglossary/g/box_cox.htm
7. Федосеев В.В. Экономико-математические методы и прикладные модели : учеб. Пособие для вузов / В.В. Федосеев [и др.]. – М. : ЮНИТИ, 2002.
8. A.Maddison, 2001. The World Economy. A Millennial Perspective, Paris, OECD. P. 264
9. The World Economy: Historical Statistics. Paris, OECD, 2003, P. 288
10. Грегори П. Экономический рост Российской империи (конец XIX – начало XX в.). Новые подсчеты и оценки. Перевод с английского И.Кузнецова и А. и Н.Тихоновых. М. Росспэн. 2003г. 256с.
11. Мельянцев В. А. Россия за три века. Указ. соч. С. 90.
12. Лященко П. И. История народного хозяйства СССР. Т. 2. М. 1956. С.406.
13. Приведение данных к нормальному распределению: преобразование Бокса-Кокса. Тематический форум. http://molbiol.ru/forums/index.php?showtopic=201368
14. Орлов А.И. О критериях согласия с параметрическим семейством http://www.newtech.ru/~orlov/kritsogl.htm
15. Мирвалиев М., Никулин М.С. / Заводская лаборатория. 1992. Т.58. № 3. С.52–58.