Теорія оптимальних рішень

УДК 519.8

В. М. ГОРБАЧУК, Г. А. ШУЛИНОК

ПРИЧИННО-СЛЕДСТВЕННАЯ СВЯЗЬ ПО ГРЕЙНДЖЕРУ

На практических примерах временных рядов предлагаются методы прогнозирования и поиска причинно-следственных связей по Грейнджеру.

Введение. При наблюдениях временных рядов часто возникает вопрос о причинно-следственных связях между ними. В отличие от кросс-секционных данных, наблюдения временных рядов упорядочены во времени: значение занятости (минимальной зарплаты, инфляции, денежной массы) в данной стране в году t может зависеть от значения этого показателя в предыдущие годы $t-1$, $t-2,\ldots,1$. Статистические свойства оценивателей обычного метода наименьших квадратов (ОМНК) как случайных переменных основаны на том, что выборки являются случайно отобранными из соответствующей генеральной совокупности [1–3]. Поскольку разные случайные выборки содержат, вообще говоря, разные значения зависимой и независимых переменных (дохода, зарплаты, стажа, семейного положения), то оцениватели ОМНК, вычисленные на разных выборках, вообще говоря, будут отличаться.

Рассмотрим в период t методы прогнозирования $y_{t+1}=\alpha_0+\alpha_1 x_{t+1}+u_{t+1}$ процесса временного ряда в будущий период $t+1$ (год, квартал, месяц, неделю, день, час, минуту) [4], которые основаны на регрессиях и исходят из информационного множества $I_t=\{y_k,x_k\}_{k=0}^{t}$, используемого в естествнном условии $E(u_t\mid I_{t-1})=0$. Обозначим f_t прогноз на шаг вперед (one-step-ahead forecast) и обозначим $e_{t+1}=y_{t+1}-f_t$ ошибку прогноза (forecast error). Так как y_{t+1} – случайная переменная, то e_{t+1} – также случайная переменная.

Самыми распространенными мерами потерь, связанных с прогнозом, являются квадрат ошибки $(e_{t+1})^2$ и модуль $|e_{t+1}|$. Значение $(e_{t+1})^2$ одинаково для ошибок $\pm e_{t+1}$; значение $|e_{t+1}|$ одинаково для ошибок $\pm e_{t+1}$.

Лучший прогноз f_t минимизирует, при данном информационном множестве I_t, ожидаемые потери

$$E[(e_{t+1})^2 \mid I_t] = E[(y_{t+1} - f_t)^2 \mid I_t].$$

Если $E(Y^2) < \infty$, а для функции $g(X)$ выполняется неравенство $E\{[g(X)]^2\} < \infty$, то по свойству условной вероятности для среднего $\mu(X)$ имеем

$$E\{[Y - \mu(X)]^2 \mid X\} \le E\{[Y - g(X)]^2 \mid X\},$$

откуда следует значение лучшего прогноза $f_t^* = E(y_{t+1} \mid I_t)$.

$f_t^* = E(y_{t+1} \mid I_t) = 0$, $t = 0,1,2,\ldots$, если $\{y_t\}_{t=0}^{\infty}$ является разностно-мартингальной последовательностью, где знание прошлого не влияет на лучший прогноз. Аналогично лучший прогноз $f_{t,h}^*$ на много шагов (h шагов) вперед (multiple-step-ahead-forecast) равен $E(y_{t+h} \mid I_t)$.

Отдачи акции часто приближают разностно-мартингальной последовательностью, но с положительным средним

$$E(y_{t+1} \mid y_t, y_{t-1}, \ldots, y_0) = E(y_{t+1}).$$

Процесс $\{y_t\}$ называют мартингалом, если

$$E(y_{t+1} \mid y_t, y_{t-1}, \ldots, y_0) = E(y_t),\ t = 0,1,2,\ldots$$

$\{\Delta y_t\}$ – разностно-мартингальная последовательность, если $\{y_t\}$ – мартингал

Методом прогнозирования также является экспоненциальное сглаживание

$$E(y_{t+1} \mid I_t) = \alpha y_t + \alpha(1-\alpha) y_{t-1} + \ldots + \alpha(1-\alpha)^{t-0} y_0 \qquad (1)$$

с параметром $\alpha \in (0,1)$ (весовой коэффициент $\alpha(1-\alpha)^{t-(t-k)} = \alpha(1-\alpha)^k$ при y_{t-k} экспонеициально убывает к 0 с ростом k). Из уравнения (1) следует

$$f_1^* = E(y_2 \mid I_1) = E(y_{1+1} \mid I_1) = \alpha y_1 + \alpha(1-\alpha) y_0,$$

$$f_2^* = E(y_3 \mid I_2) = E(y_{2+1} \mid I_2) = \alpha y_2 + \alpha(1-\alpha) y_1 + \alpha(1-\alpha)^2 y_0 =$$

$$= \alpha y_2 + (1-\alpha)[\alpha y_1 + \alpha(1-\alpha) y_0] = \alpha y_2 + (1-\alpha) E(y_2 \mid I_1) = \alpha y_2 + (1-\alpha) f_1^*,$$

и т. д. Отсюда в предположении $y_0 = f_0^*$ получаем рекуррентную зависимость

$$f_t^* = E(y_{t+1} \mid I_t) = \alpha y_t + \alpha(1-\alpha) f_{t-1}^*,\ t = 1,2,\ldots,$$

которая требует выбора параметра α. Регрессионные методы позволяют оценивать значения таких параметров.

Если для прогнозирования применять статическую регрессионную модель

$$y_t = \beta_0 + \beta_1 z_t + u_t \qquad (2)$$

с единственной объясняющей переменной z_t, а в период t известны параметры β_0, β_1, то лучший прогноз (при условии знания z_{t+1}) определяется

$$f_t^* = E(y_{t+1} \mid z_{t+1}, y_t, z_t, \ldots, y_1, z_1) = \beta_0 + \beta_1 E(z_{t+1} \mid I_t) + E(u_{t+1} \mid I_t).$$

Если объясняющая переменная не является временным трендом или сезонной переменной, то значение z_{t+1} в будущий период $(t+1)$ неизвестно. Кроме того, если $\{u_t\}$ содержит серийную корреляцию, то $E(u_{t+1} \mid I_t) \neq 0$.

Для прогнозирования более подходящей является не статическая модель (2), а модель с только лаговыми значениями y и z:

$$y_t = \delta_0 + \alpha_1 y_{t-1} + \gamma_1 z_{t-1} + u_t, \qquad (3)$$
$$E(u_t \mid I_{t-1}) = 0.$$

Отсюда, используя $E(y_{t-1} \mid I_{t-1}) = y_{t-1}$, $E(z_{t-1} \mid I_{t-1}) = z_{t-1}$, получаем

$$y_{t+1} = \delta_0 + \alpha_1 y_t + \gamma_1 z_t + u_{t+1},$$
$$f_t^* = E(y_{t+1} \mid I_t) = \delta_0 + \alpha_1 E(y_t \mid I_t) + \gamma_1 E(z_t \mid I_t) + E(u_{t+1} \mid I_t) =$$
$$= \delta_0 + \alpha_1 y_t + \gamma_1 z_t.$$

Можно по выборке размера n найти ОМНК-оценки $\hat{\delta}_0$, $\hat{\alpha}_1$, $\hat{\gamma}_1$ для значений δ_0, α_1, γ_1 параметров зависимости (3) и построить точечный прогноз (point forecast) для y_{n+1}

$$\hat{f}_n = \hat{\delta}_0 + \hat{\alpha}_1 y_t + \hat{\gamma}_1 z_t, \qquad (4)$$

а в период $(n+1)$ вычислить ошибку прогноза

$$\hat{e}_{n+1} = y_{n+1} - \hat{f}_n. \qquad (5)$$

Интервал прогноза (forecast interval) определяют таким самым способом, как интервал предвидения (prediction interval) 95 % при классических предположениях линейной модели. Зависимость (3) содержит лаговую зависящую переменную y_{t-1} и поэтому не удовлетворяет этим предположениям, но интервал прогноза остается приближенно валидным, если погрешность $(u_t \mid I_{t-1})$ (погрешность u_t при данной информационном множестве I_{t-1}) является нормально распределенной с нулевым средним и постоянной дисперсией. При такой погрешности $(u_t \mid I_{t-1})$, ОМНК-оценки являются приближенно нормально распределенными с обычными ОМНК-дисперсиями, а погрешность u_{t+1} не зависит от ОМНК-оценивателей, имеет нулевое среднее и дисперсию σ^2.

Значения $\hat{f}_n$ точечного прогноза и его стандартной ошибки (standard error) $SE(\hat{f}_n)$ можно получить как пересечение (intercept) и его стандартную ошибку

для регрессии y_t по $(y_{t-1} - y_n)$ и $(z_{t-1} - z_n)$, $t = 1, 2, \ldots, n$. Из уравнения (5) следует

$$[SE(\hat{e}_{n+1})]^2 = [SE(\hat{f}_n)]^2 + \hat{\sigma}^2, \tag{6}$$

где $\hat{\sigma}$ – оценка для σ. Тогда интервал прогноза 95 % задается

$$(\hat{f}_n - 1.96\,SE(\hat{e}_{n+1}), \hat{f}_n + 1.96\,SE(\hat{e}_{n+1})). \tag{7}$$

Ошибка $SE(\hat{f}_n)$ приближенно пропорциональна $\frac{1}{\sqrt{n}}$ и поэтому мала по сравнению с $\hat{\sigma}$ (мерой неопределенности погрешности u_{n+1}).

На данных США 1948–1996 гг. табл. 1 [5] (с помощью MS Excel) оценим параметры простой авторегрессии AR(1)

$$U_t = \beta_0 + \beta_1 U_{t-1} + u_t$$

для уровня U_t (%) гражданской безработицы (civilian unemployment) в году t:

$$\hat{U}_t = \underset{(0.577)}{1.572} + \underset{(0.097)}{0.732}\ U_{t-1}, \tag{8}$$

Здесь выражение в круглых скобках означает стандартную ошибку соответствующей оценки параметра; $n = 1996 - 1948 = 48$; $R^2 = 0.554$; нормированная (adjusted) величина $\overline{R}^2 = 0.544$; $\hat{\sigma} = 1.049$.

На этих же данных оценим параметры модели

$$U_t = \beta_0 + \beta_1 U_{t-1} + \alpha_1 P_{t-1} + v_t,$$

обобщающей зависимость (8) путем учета инфляции – роста P_{t-1} (%) средних потребительских цен (consumer price index, CPI) в году $(t-1)$:

$$\hat{U}_t = \underset{(0.490)}{1.304} + \underset{(0.084)}{0.647}\ U_{t-1} + \underset{(0.041)}{0.184}\, P_{t-1}, \tag{9}$$

$n = 48$; $R^2 = 0.691$; нормированная величина $\overline{R}^2 = 0.677$; $\hat{\sigma} = 0.883$.

Хотя модель (9) имеет лучшее значение нормированной величины $\overline{R}^2$, чем модель (8), и довольно значимую t-статистику для инфляции ($4.46 >> 1.96$), это не обязательно означает, что модель (9) дает лучший прогноз безработицы на следующий после наблюдений 1997 г., чем модель (8). По модели (9)

$\hat{U}_{1997} = 1.304 + 0.647\,U_{1996} + 0.184\,P_{1996} = 1.304 + 0.647 \times 5.4 + 0.184 \times 3.0 = 5.35,$

а по модели (8) –

$$\hat{U}_{1997} = 1.572 + 0.732\ U_{1996} = 1.572 + 0.732 \times 5.4 = 5.53.$$

Наблюдение дает $U_{1997} = 4.9$ (табл. 2 [5]), что ближе к прогнозу модели (9).

На этих же данных оценим параметры модели

$$U_t = \beta_0 + \beta_1 (U_{t-1} - 5.4) + \alpha_1 (P_{t-1} - 3.0) + w_t:$$

$$\hat{U}_t = \underset{(0.137)}{5.348} + \underset{(0.084)}{0.647}\ (U_{t-1} - 5.4) + \underset{(0.041)}{0.184}\ (P_{t-1} - 3.0); \tag{10}$$

$n = 48$; $R^2 = 0.691$; нормированная величина $\overline{R}^2 = 0.677$; $\hat{\sigma} = 0.883$. Тогда из соотношений (4), (10) следует $SE(\hat{f}_n) = 0.137$, откуда в силу равенства (6) имеем

$$SE(\hat{\varepsilon}_{n+1}) = \{[SE(\hat{f}_n)]^2 + \hat{\sigma}^2\}^{0.5} = (0.137^2 + 0.883^2)^{0.5} = 0.893.$$

Таким образом, интервал прогноза по модели (9) определяется формулой (7):

$$(\hat{f}_n - 1.96\,SE(\hat{\varepsilon}_{n+1}), \hat{f}_n + 1.96\,SE(\hat{\varepsilon}_{n+1})) =$$
$$= (5.348 - 1.96 \times 0.893,\ 5.348 + 1.96 \times 0.893) = (3.597,\ 7.100).$$

Очевидно, $U_{1997} = 4.9$ принадлежит этому интервалу.

Обычно требуется вырабатывать прогноз для каждого периода (времени): например, в 1996-м году требуется выработать прогноз U_{1997}; когда становятся известными значения U_{1997}, P_{1997}, требуется выработать прогноз U_{1998}. Если вырабатывать прогноз U_{1998} на основе модели (3), то это можно делать по крайней мере двумя способами:

1) использовать формулу (4) $\hat{f}_{1998} = \hat{\delta}_{96} + \hat{\alpha}_{96} U_{1997} + \hat{\gamma}_{96} P_{1997}$, где параметры $\hat{\delta}_{96}$, $\hat{\alpha}_{96}$, $\hat{\gamma}_{96}$ оцениваются на 49 наблюдениях $\{U_t, P_t\}_{t=1948}^{1996}$;

2) использовать формулу (4) $\hat{f}_{1998} = \hat{\delta}_{97} + \hat{\alpha}_{97} U_{1997} + \hat{\gamma}_{97} P_{1997}$, где параметры $\hat{\delta}_{97}$, $\hat{\alpha}_{97}$, $\hat{\gamma}_{97}$ оцениваются на 50 наблюдениях $\{U_t, P_t\}_{t=1948}^{1997}$, включающих предыдущие 49 наблюдений $\{U_t, P_t\}_{t=1948}^{1996}$.

Для способа 1) применим соотношение (9) и табл. 2 [5]:

$\hat{f}_{1998} = \hat{\delta}_{96} + \hat{\alpha}_{96} U_{1997} + \hat{\gamma}_{96} P_{1997} = 1.304 + 0.647 \times 4.9 + 0.184 \times 2.3 = 4.90.$

Для способа 2) на данных США 1948–1997 гг. табл. 1, 2 [5] оценим параметры авторегрессии

$$U_t = \beta_0 + \beta_1 U_{t-1} + u_t:$$
$$\hat{U}_t = \underset{(0.572)}{1.549} + \underset{(0.096)}{0.734}\ U_{t-1}; \qquad (11)$$

$n = 1997 - 1948 = 49$; $R^2 = 0.554$; $\overline{R}^2 = 0.544$; $\hat{\sigma} = 1.041$. На этих же данных оценим параметры модели

$$U_t = \beta_0 + \beta_1 U_{t-1} + \alpha_1 P_{t-1} + v_t:$$
$$\hat{U}_t = \underset{(0.484)}{1.286} + \underset{(0.083)}{0.648}\ U_{t-1} + \underset{(0.041)}{0.185}\ P_{t-1}, \qquad (12)$$

$n = 49$; $R^2 = 0.691$; $\overline{R}^2 = 0.677$; $\hat{\sigma} = 0.876$. Оценки (11) и (12) близки к оценкам (8) и (9) соответственно, так как учитывают одно дополнительное наблюдение.

По модели (12)

$\hat{U}_{1998} = 1.286 + 0.648\,U_{1997} + 0.185\,P_{1997} = 1.286 + 0.648 \times 4.9 + 0.185 \times 2.3 = 4.88,$

а по модели (11) –

$$\hat{U}_{1998} = 1.549 + 0.734\ U_{1997} = 1.549 + 0.734 \times 4.9 = 5.15.$$

Наблюдение дает $U_{1998} = 4.5$ (табл. 2 [5]), что незначительно (на 0.02) ближе ближе к прогнозу модели (12) способа 2), чем модели (9) способа 1).

Зависимость (3) является одним из уравнений векторной авторегрессионной (vector autoregressive, VAR) модели. Если авторегрессионная (AR) модель касается единственого временного ряда $\{y_t\}$, где переменная y_t выражается через собственное прошлое y_{t-1}, $y_{t-2,}\ldots$, y_0, то векторная авторегрессионная модель касается по крайней мере двух рядов $\{y_t\}$ и $\{z_t\}$, где каждая переменная y_t (или z_t) выражается как через собствнное прошлое y_{t-1}, $y_{t-2,}\ldots$, y_0, так и через прошлое другой переменной z_{t-1}, $z_{t-2,}\ldots$, z_0:

$$y_t = u_t + \delta_0 + \alpha_1 y_{t-1} + \gamma_1 z_{t-1} + \alpha_2 y_{t-2} + \gamma_2 z_{t-2} + \cdots, \tag{13}$$

$$z_t = w_t + \eta_0 + \beta_1 y_{t-1} + \rho_1 z_{t-1} + \beta_2 y_{t-2} + \rho_2 z_{t-2} + \cdots, \tag{14}$$

$$E(u_t \mid I_{t-1}) = 0 = E(w_t \mid I_{t-1}).$$

Выбор числа лагов в зависимостях (13) и (14) виден на примере модели (9): значение F-теста на совместную значимость переменных U_{t-2} и P_{t-2} подтверждает, что только одного лага для безработицы и инфляции достаточно. Если период – это год, то 1–2 лага обычно достаточно; если период – квартал, то требуется 4–8 лагов; если период – месяц, то требуется от 6 до 24 лагов.

Уравнения VAR (13) и (14) полезны при прогнозировании переменной y, для чего требуется оценить и проанализировать уравнение (13). Модель VAR может включать 3 ряда $\{y_t\}$, $\{z_t\}$, $\{x_t\}$. В предположении гомоскедастичности можно применять ОМНК для оценки параметров модели.

Говорят, что z влияет по Грейнджеру (Granger causes) на y, если

$$E(y_t \mid z_{t-1}, y_{t-1}, z_{t-2}, y_{t-2}, \ldots, z_0, y_0) \neq E(y_t \mid y_{t-1}, y_{t-2}, \ldots, y_0);$$

неравенство не означает одновременной (contemporaneous) причинности между z и y, и нельзя сказать, является ли z экзогенной или эндогенной переменной в зависимостях (13) и (14). Поэтому причинность по Грейнджеру не применяют для кросс-секционных данных (Грейнджер – Нобелевский лауреат 2002 г.) [6, 7].

Говорят, что z влияет по Грейнджеру на y при условии g, если

$$E(y_t \mid z_{t-1}, y_{t-1}, z_{t-2}, y_{t-2}, \ldots, z_0, y_0) \neq E(y_t \mid y_{t-1}, g_{t-1,} y_{t-2}, g_{t-2}, \ldots, y_0, g_0).$$

Может быть, что z влияет по Грейнджеру на y, но z не влияет по Грейнджеру на y при условии g. Если z – рост предложения денег, y – рост реального валового внутреннего продукта, g – изменение процентных ставок, то причинно-следственная связь по Грейнджеру имеет прикладное значение [8, 9].

Предположим, y_t определяют 3 лаговые переменные y_{t-1}, y_{t-2}, y_{t-3}:

$$y_t = \delta_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \alpha_3 y_{t-3} + u_t,$$

$$E(u_t \mid y_{t-1}, y_{t-2}, \ldots, y_0) = 0.$$

Гипотеза H_0 о том, что переменная z_{t-1} не влияет по Грейнджеру на y_t, сводится к t-тесту для z_{t-1} в зависимости

$$y_t = \delta_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \alpha_3 y_{t-3} + \gamma_1 z_{t-1} + u_t;$$

гипотеза H_0 о том, что переменные z_{t-1}, z_{t-2} не влияют по Грейнджеру на y_t, сводится к F-тесту на совместную значимость для z_{t-1} и z_{t-2} в зависимости

$$y_t = \delta_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \alpha_3 y_{t-3} + \gamma_1 z_{t-1} + \gamma_2 z_{t-2} + u_t.$$

В случае гетероскедастичности можно использовать робастные варианты тестов. При гипотезе H_0 серийная корреляция отсутствует, поскольку модель является динамически полной.

Перед построением модели (9) была построена модель (8), где число лагов равно 1 (модель (8) принадлежит классу AR(1)). Модель (9) говорит, что инфляция влияет по Грейнджеру на безработицу.

В. М. Горбачук,Г. О. Шулінок

ПРИЧИННО-НАСЛІДКОВИЙ ЗВ'ЯЗОК ЗА ГРЕЙНДЖЕРОМ

На практичних прикладах часових рядів пропонуються методи прогнозування та пошуку причинно-наслідкових зв'язків за Грейнджером.

V.M. Gorbachuk, G.O. Shulinok

GRANGER CAUSALITY

The methods of forecasting and Granger causality search are proposed, based upon practical examples of time series.

1. *Горбачук В. М.* Эконометричне програмування TSP та EViews. Препр. 96-14. – К.: Ін-т кібернетики імені В. М. Глушкова НАН України, 1996. – 24 с.
2. *Горбачук В. М.* Макроекономічні методи. – К.: Альтерпрес, 1999. – 263 с.
3. *Wooldridge J. M.* Introductory econometrics: a modern approach. 4-th edition. – Mason, OH: Cencage Learning, 2009. – 865 p.
4. *Diebold F. X.* Elements of forecasting. – Cincinnati, OH: South-Western, 1998.
5. *Горбачук В. М., Кривонос Ю. Г.* Особенности регрессионного анализа временных рядов // Компьютерная математика. – 2012. – № 2. – С. 3–12.
6. *Gorbachuk V.* Causality in time series analysis // Nonlinear analysis and applications. – Kyiv: NTUU "KPI", 2012. – P. 30.
7. *Gorbachuk V. M.* Regression analysis of time series and Granger causality // 14-та міжнародна наукова конференція імені академіка М. Кравчука. Т. 3. – К.: НТУУ „КПІ", 2012. – С. 11.
8. *Stock J. H., Watson M. W.* Interpreting the evidence on money-income causality // Journal of econometrics. – 1989. – V. 40. – P. 161–181.
9. *Горбачук В. М.* Методи індустріальної організації. – К.: А. С. К., 2010. – 224 с.

Получено 22.05.2012