

DYNAMICZNE MODELE EKONOMETRYCZNE

VIII Ogólnopolskie Seminarium Naukowe, 9-11 września 2003 w Toruniu
Katedra Ekonometrii i Statystyki, Uniwersytet Mikołaja Kopernika w Toruniu

Anna Szmit
Poitechnika Łódzka

Analiza jakości prognoz w zależności od długości horyzontu prognozy

1. Wstęp

Intencją każdego prognosty jest uzyskanie możliwie najdokładniejszych prognoz, jednak jak wiadomo proces predykcji jest związany nieodłącznie z generowaniem błędów, wynikających z różnych przyczyn. Jest też rzeczą naturalną, że dokładność prognoz wykonywanych na pewien okres do przodu pogarsza się w miarę wydłużania się horyzontu prognozy. Ten fakt ma również kilka przyczyn.

W prezentowanym referacie zaprezentowano próbę analizy tempa pogarszania się dokładności prognoz w miarę wydłużania się horyzontu prognozy na przykładzie prognoz generowanych przez wybrane jednorównaniowe modele regresji liniowej.

Podstawowymi metodami i miernikami oceny dokładności i trafności prognoz są błędy średnie, zarówno szacowane z góry – *ex ante*, jak i znane dopiero po wygaśnięciu prognoz błędy *ex post*. Analiza jakości prognoz w funkcji długości horyzontu prognozy ma zatem dwa, ściśle związane ze sobą wątki.

2. Opis problemu

Wśród przyczyn błędów prognoz na szczególne podkreślenie zasługują:

- niewłaściwa postać modelu,
- złe wartości zmiennych objaśniających dla okresu prognozy,
- losowa zmienność zmiennej prognozowanej.

Pierwszej z wymienionych przyczyn należy się oczywiście wystrzegać w procesie wyboru odpowiedniej metody prognozowania i oceny jakości modelu.

Jednak nawet model, który w pewnym okresie bardzo dobrze odwzorowuje badane zjawisko, po jakimś czasie przestaje być adekwatną metodą opisu. Jest to w szczególności związane z długością horyzontu prognozy. Na ogół wyróżnia się¹ krótki horyzont prognoz jako taki, w którym zachodzą tylko zmiany ilościowe zjawiska, przy zachowaniu zależności jakościowych, natomiast w długim horyzoncie zachodzą zmiany jakościowe, czyli na przykład zmiana postaci istniejących między zależnościami przyczynowo-skutkowymi i idąca za nią konieczność zmiany postaci modelu opisowego. Stąd model odpowiedni w krótkim horyzoncie czasowym na ogół nie jest właściwy dla długiego okresu.

Źródłem błędów bywa także nieznanie wartości zmiennych objaśniających dla okresu prognozy. Aby w ogóle wygenerować prognozę w takim przypadku przy pomocy opisowego modelu ekonometrycznego, konieczna jest przynajmniej znajomość prognoz wartości tych zmiennych. Jednak jeśli ich prognozy są obciążone błędem, to błąd ten będzie zwiększał² wielkość błędu prognozy modelu.

Losowa zmienność jest właściwa większości zjawisk. Wielkość tej zmienności daje się szacować na podstawie obserwacji szeregu danych w procesie estymacji parametrów modelu przy pomocy wariancji resztowej.

O ocenie metody prognozowania, w szczególności modelu ekonometrycznego, decyduje bardzo często praktyczna przydatność generowanych prognoz, to znaczy najczęściej możliwie mała wielkość uzyskiwanych błędów. Szczególne znaczenie ma zatem wielkość średniego³ błędu prognozy *ex post*, na przykład w postaci:

$$S(y_T^P) = \sqrt{\frac{\sum_{\tau=1}^T (y_\tau^P - \hat{y}_\tau^P)^2}{T}}, \quad (1)$$

gdzie:

y_τ^P - wartość rzeczywista zmiennej objaśnianej dla momentu τ prognozy, $\tau = 1, \dots, T$,

\hat{y}_τ^P - wartość teoretyczna zmiennej objaśnianej (generowana przez model) dla momentu τ prognozy.

Często stosowanymi miernikami są także błędy procentowe, np. średnia z wartości bezwzględnych błędów procentowych *MAPE*:

¹ Porównaj Cieślak, (1999), s. 24. W pracy Zeliaś (1997), s. 23 przedstawiono umowny podział, zgodnie z którym za krótkie uważa się prognozy do roku, średnioterminowe obejmują okres 2-5 lat, a długoterminowe prognozy są wykonywane na więcej niż 5 lat.

² Porównaj Zeliaś (1997), s. 68.

³ W niektórych przypadkach bardziej istotna jest wielkość błędu maksymalnego.

$$MAPE = \frac{1}{T} \sum_{\tau=1}^T \left| \frac{y_{\tau}^P - \hat{y}_{\tau}^P}{y_{\tau}^P} \right|. \quad (2)$$

Powyższa wielkość jest jednak możliwa do wyznaczenia dopiero w warunkach, gdy znane są wartości zmiennej prognozowanej y dla okresu prognozy. Aby ocenić i przewidzieć z góry wielkość błędu prognozy dla różnych horyzontów czasowych, potrzebne jest oszacowanie wielkości błędu predykcji *ex ante*. Dla jednorównaniowego modelu regresji liniowej, dla którego spełnione są założenia prognostyczne⁴, wartość oczekiwana błędu predykcji wynosi zero, zaś ocena średniego błędu predykcji, czyli pierwiastka z wariancji błędu predykcji⁵ wynosi

$$S(\hat{y}_{\tau}^P) = \sqrt{S_e^2 (x_{\tau}^P (X^T X)^{-1} (x_{\tau}^P)^T + 1)}, \quad (3)$$

gdzie

$X = [x_{ij}]_{\substack{j=1, \dots, n \\ i=0, 1, \dots, k}}$ - macierz n obserwacji k zmiennych objaśniających w mo-

delu regresji liniowej ($x_{i0} = 1$),

$x_{\tau}^P = [1 \quad x_{1\tau}^P \quad \dots \quad x_{k\tau}^P]$ - wektor wartości zmiennych objaśniających na mo-

ment τ prognozy,

S_e^2 - wariancja reszt modelu.

Stosuje się również względny błąd średni predykcji w postaci:

$$V_{\tau} = \frac{S(\hat{y}_{\tau}^P)}{\hat{y}_{\tau}^P}. \quad (4)$$

Ze wzoru (3) wynika, że średni błąd predykcji jest tym większy, im dalej wektor x_{τ}^P jest od wektora \bar{x} . W szczególności, pochodne cząstkowe $S(\hat{y}_{\tau}^P)$ po zmiennej x_i wyrażają się wzorem

$$\frac{dS(\hat{y}_{\tau}^P)}{dx_i} = S_e \frac{b_{ii}x_i + \sum_{\substack{j=0..k \\ j \neq i}} b_{ij}x_j}{\sqrt{b_{ii}x_i^2 + 2 \sum_{\substack{j=0..k \\ j \neq i}} b_{ij}x_j + \sum_{\substack{j=0..k \\ j \neq i}} \sum_{\substack{l=0..k \\ l \neq i}} b_{jl}x_jx_l + 1}}$$

gdzie b_{ij} są elementami macierzy $(X^T X)^{-1}$.

Zatem np. w przypadku $k=1$

⁴ Porównaj np. por. Zeliaś (1997), s. 50-52, Cieślak (1999), s. 115-116, Witkowska (2002), s. 260-261.

⁵ W przypadku, kiedy w modelu stosuje się oceny estymatorów jego parametrów (zamiast nieznanymi wartościami parametrów), nie uwzględniając stochastycznej natury wektora x_{τ}^P .

$$\frac{dS(\hat{y}_\tau^P)}{dx} = \frac{S_e}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \cdot \frac{x - \bar{x}}{\sqrt{(x - \bar{x})^2 + \left(1 + \frac{1}{n}\right) \sum_{i=1}^n (x_i - \bar{x})^2}}$$

oraz

$$\frac{d^2 S(\hat{y}_\tau^P)}{dx^2} = \frac{S_e \left(1 + \frac{1}{n}\right) \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}{\left((x - \bar{x})^2 + \left(1 + \frac{1}{n}\right) \sum_{i=1}^n (x_i - \bar{x})^2\right)^{3/2}},$$

czyli $S(\hat{y}_\tau^P)$ jest funkcją wypukłą względem x , co oznacza, że w miarę oddalania się od \bar{x} średni błąd predykcji rośnie coraz szybciej.

Powyższe uwagi nie oznaczają jednak automatycznego zwiększania się błędu predykcji w miarę wydłużania horyzontu prognozy dla każdego modelu regresji liniowej. Istnieje wszak wiele modeli, dla których wartości zmiennych objaśniających nie mają charakteru monotonicznego w okresie prognozy, w szczególności modele bez takich zmiennych jak zmienna czasowa t . Jednym z przykładów może być model sezonowy bez trendu o macierzy zmiennych objaśniających (powtarzających się cyklicznie również w okresie prognozy) postaci podobnej do przedstawionej poniżej:

$$X = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

W takim przypadku średni błąd predykcji jest stały w całym okresie prognozy. Przykład: modele ze zmiennymi opóźnionymi i bez.

Zmienną objaśnianą (i prognozowaną) w prezentowanych badaniach był szereg zużycia energii elektrycznej o częstotliwości pomiaru jedna godzina. Poniżej zostaną zaprezentowane wyniki analizy wielkości błędów prognozy dla dwóch modeli⁶ regresji liniowej. W obu modelach jako zmiennych objaśniających użyto wielu⁷ zmiennych zero-jedynkowych reprezentujących warianty zmiennych jakościowych opisujących bieżący moment w czasie i znanych z

⁶ Dokładny opis modeli znajduje się w pracy Szmit, (2002).

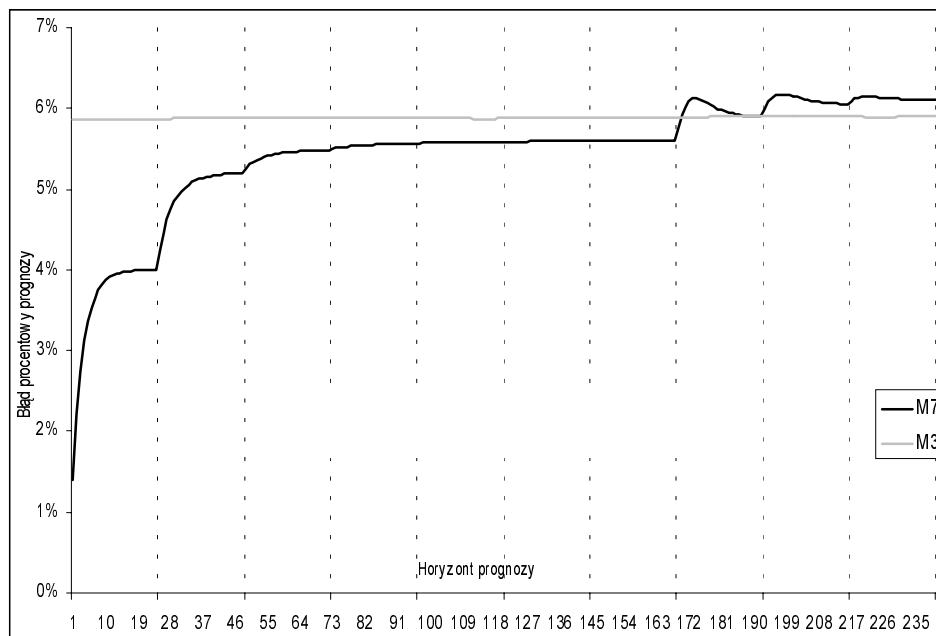
⁷ Obydwa modele zawierały ponad sto zmiennych objaśniających.

dowolnie dużym wyprzedzeniem. W pierwszym z omawianych modeli (zwanym M3) były to jedyne zmienne objaśniające modelu. W drugim z prezentowanych, modelu M7 oprócz wymienionych zmiennych użyto także bloku zmiennych opóźnionych reprezentujących przeszłe wartości zmiennej objaśnianej z opóźnieniem o 1 (1 godzinę), 2, 3, 24 (dobę), 25, 26 oraz 168 (tydzień), 169, 170. Użyte modele nie zawierały składnika trendu i wielkość wyrażenia⁸ (3) nie zmieniała się istotnie w okresie prognozy. W przypadku modelu M3 wielkość ta reprezentowała oszacowanie średniego błędu predykcji, zaś dla modelu M7 była średnim błędem predykcji tylko dla prognozy o horyzoncie 1, ponieważ dla większych wyprzedzeń czasowych nie były znane wielkości rzeczywiste wszystkich zmiennych objaśniających – tzn. nie były znane pewne wartości zmiennych opóźnionych, a tylko ich wartości prognozowane (prognozy uzyskano przy pomocy tego samego modelu M7). Zatem wielkość średniego błędu predykcji modelu M7 rośnie wraz z wydłużaniem się horyzontu prognozy. Należy oczekiwać, że wzrost błędu prognozy powinien być widoczny także dla błędów *ex post*. I tak jest rzeczywiście, co można zaobserwować na rysunku 1.

Prognozy były wykonywane dla horyzontu od 1 do 8000 (do jedenastu miesięcy), jednak na rysunku uwidoczniono, dla lepszej czytelności, tylko pierwsze 240 wyprzedzeń. Wszystkie błędy na nim widoczne są średnimi błędami prognoz na τ -tą godzinę prognozy (dla prognozy o horyzoncie τ , $\tau=1, \dots, 240$). Można zauważyć, że błędy prognoz modelu M3 bez zmiennych opóźnionych praktycznie są stałe. Natomiast w modelu M7 ze zmiennymi opóźnionymi bardzo dobrze widoczne są momenty dołączania kolejnych prognoz zmiennych opóźnionych zamiast wartości oryginalnych. Wartości błędu dla takich opóźnień zaczynają zdecydowanie rosnać, na początku szybko, a wraz ze wzrostem opóźnienia poziom błędu się stabilizuje. Można zatem stwierdzić, że główną przyczyną zwiększania się błędu prognozy w miarę wydłużania horyzontu w tego rodzaju modelach jest niedokładna wartość zmiennej objaśnianej. Można zauważyć także, że model M7 nie powinien być stosowany do prognoz o horyzoncie dłuższym niż 7 dni. Po tym czasie jakość modelu obniża się znacznie i mniejsze wartości błędów prognoz uzyskuje się przy pomocy modelu M3.

Horyzont prognoz generowanych w opisywanych badaniach (do 11 miesięcy) należy do zakresu horyzontu krótkiego, to znaczy nie ujawniają się w istotny sposób zmiany jakościowe opisywanego zjawiska. Zużycie energii elektrycznej należy do zjawisk o dość dużej regularności i dużej bezwładności szeregu czasowego.

⁸ Wartości błędów predykcji nie były szacowane ze względu na problem obliczeniowy: macierz X miała 27048 wierszy i ponad 100 kolumn w przypadku obu modeli.



Rysunek 1. Wykres wartości średniego błędu prognozy (MAPE) na moment τ prognozy dla modeli M3 M7. Źródło: opracowanie własne.

Literatura

- Cieślak, M. (red.), (1999), *Prognozowanie gospodarcze. Metody i zastosowania*, Wydawnictwo Naukowe PWN, Warszawa.
- Szmit, A. (2002), *Prognozowanie zapotrzebowania na energię elektryczną. Studium empiryczne dla regionu łódzkiego*, Rozprawa doktorska przygotowana pod kierunkiem prof. dr hab. D. Witkowskiej, Politechnika Łódzka, Łódź (maszynopis).
- Witkowska, D. (2002), *Sztuczne sieci neuronowe i metody statystyczne*. Wydawnictwo C. H. Beck, Warszawa.
- Zeliaś, A. (1997), *Teoria prognozy*, Polskie Wydawnictwo Ekonomiczne, Warszawa.