Статистические методы оценки качества прогнозов экономической деятельности

Оценка качества прогнозов: простейшие методы

Аннотация

Статья посвящена анализу качества прогнозирования показателей экономической деятельности. Рассмотрены простейшие статистические методы, при помощи которых можно сравнить несколько прогнозов конкретного показателя, и набор тестов, позволяющих ответить на вопрос о статистической значимости разницы между этими прогнозами.

Ключевые слова: прогнозирование, оценка качества прогнозов, экономическая деятельность, экономические показатели

Турунцева Марина Юрьевна
ст. научный сотрудник, Институт экономической политики им. Е.Т. Гайдара, и. о. зав. лабораторией краткосрочного прогнозирования
m.turuntseva@gmail.com

Прогнозирование показателей экономической деятельности является неотъемлемой составляющей экономического процесса. Существует много методов прогнозирования, в частности, экспертные оценки, обследования потребителей и предпринимателей, экстраполирование, модели временных рядов, эконометрические системы. В этой связи встает вопрос об оценке качества прогнозов, полученных различными способами. Существует стандартный набор простейших статистик качества прогнозов и ряд довольно простых тестов, позволяющих ответить на вопрос о значимости различий между прогнозами того или иного показателя, если их было несколько. Необходимо отметить, что предложенные методы не зависят от того, каким из перечисленных выше способов получены прогнозы.

Простейшие статистики качества прогнозов

К простейшим статистикам качества прогнозов относятся средняя абсолютная процентная ошибка прогнозирования (*Mean Absolute Percent Error – MAPE*), средняя абсолютная ошибка прогнозирова-

ния (*Mean Absolute Error – MAE*), корень квадратный из средней квадратичной ошибки прогнозирования (*Root Mean Squared Error – RMSE*)[1].

Средняя абсолютная процентная ошибка прогнозирования является абсолютной мерой качества прогнозов в том смысле, что позволяет оценить его независимо от других прогнозов: достаточно выбрать некий уровень средней ошибки (например, 5%) и сравнивать рассчитанное по статистике значение с этим тестовым уровнем. Если расчетное значение меньше тестового, то прогноз считается хорошим, если больше – плохим.

Две другие меры качества прогнозов (*MAE* и *RMSE*) являются относительными, то есть могут быть использованы для сравнения двух (или более) различных прогнозов одного и того же показателя между собой: лучшим считается тот прогноз, у которого значение *MAE* или *RMSE* меньше. При этом, очевидно, этот лучший прогноз может быть хорошим или плохим с точки зрения *MAPE*. Обычно все эти статистики не противоречат друг другу, то есть выбирают в качестве лучшего один и тот же прогноз, но наиболее часто для сравнения прогнозов используется *RMSE*.

Главными достоинствами всех перечисленных выше статистик является простота их расчета и независимость от свойств ошибок прогнозирования, главным недостатком – то, что они не позволяют получить ответ на вопрос о том, являются ли два прогноза показателя разными со статистической точки зрения. Поясним, что мы имеем в виду.

Пусть у нас есть два различных прогноза одного и того же показателя. Например, эти прогнозы получены по двум разным моделям *A* и *B*. Мы знаем, что модель *A* является довольно простой (например, с точки зрения методов ее оценки), а модель *B*, напротив – сложной. Рассчитав простейшие статистики качества получаем, что модель *B* обладает чуть лучшими характеристиками, например, ее *MAPE* равно 4,9%, а *MAPE* модели *A* – 5,3%. На первый взгляд, модель *B* лучше. Но мы знаем, что для ее оценки требуется гораздо больше усилий по сравнению с моделью *A*. Соответственно, возни-

средняя абсолютная процентная ошибка прогнозирования является абсолютной мерой качества прогнозов в том смысле, что позволяет оценить его независимо от других прогнозов

[1] Здесь и далее все необходимые для расчетов статистик формулы приведены в Приложении.

кает вопрос: а стоит ли тратить много усилий на оценку модели, прогноз по которой получается не намного лучше, чем при использовании с гораздо более простой модели?

Для ответа на этот вопрос можно использовать специальные тесты: F-тест, тест Моргана – Грейнджера – Ньюболда, тест Миза – Рогоффа, тест знаков и ранговый тест знаков Вилкоксона, которые позволяют выяснить (проверить гипотезу) являются ли множества прогнозов, полученных двумя разными способами, различными с формальной (статистической) точки зрения. Отметим актуальность проблемы: многие эмпирические исследования показывают, что прогнозы по простым моделям очень часто оказываются лучше с точки зрения простейших статистик качества, чем прогнозы, полученные более сложными методами.

Тесты для проверки гипотезы о совпадении прогнозов

Самым простым способом проверки гипотезы о совпадении прогнозов, полученными двумя различными способами (*A* и *B*), является *F*-тест, который рассчитывается как отношение выборочной ковариации между ошибками прогнозирования, полученными по различным моделям, к выборочной дисперсии ошибки прогнозирования, полученной по модели *B*. Для возможности применения теста необходимо, чтобы ошибки прогнозирования удовлетворяли всем стандартным требованиям, то есть имели нулевой средний уровень, являлись нормальными, а также серийно и одновременно некоррелированными. Такие серьезные ограничения являются главным недостатком теста, поскольку сильно ограничивают возможности его корректного использования применительно к реальным данным.

В тесте Моргана – Грейнджера – Ньюболда можно ослабить предположение об одновременной коррелированности ошибок прогнозирования. Более того, Диболд и Мариано показали, что единственным предположением, которое не может быть ослаблено по сравнению с *F*-тестом, является необходимость использования в качестве базовой статистики *RMSE*.

...меры качества прогнозов (MAE и RMSE) являются относительными, то есть могут быть использованы для сравнения двух (или более) различных прогнозов одного и того же показателя

Тест Миза – Рогоффа можно использовать, если ошибки прогнозирования являются и серийно, и одновременно коррелированными. И он совпадает с тестом Моргана – Грейнджера – Ньюболда, если ряды ошибок не являются серийно коррелированными.

Диболд и Мариано предложили тест, являющийся устойчивым к различным отклонениям от стандартных предположений о свойствах ошибок прогнозирования – они ослабили все предположения классического *F*-теста. С этой точки зрения, данный тест является универсальным инструментом проверки гипотезы об отсутствии значимых различий между прогнозами.

Главным недостатком всех рассмотренных тестов, является то, что они дают хорошие результаты, если в наличии имеются длинные ряды прогнозов. Но чаще всего это условие не выполняется. В таком случае можно использовать тест знаков и ранговый тест знаков Вилкоксона, который является более мощным тестом (то есть при прочих равных условиях дает более достоверные результаты) по сравнению с тестом знаков.

...многие эмпирические исследования показывают, что прогнозы по простым моделям очень часто оказываются лучше.., чем прогнозы, полученные более сложными методами

Вывод

Все рассмотренные тесты хорошо работают при большом количестве наблюдений и при выполнении необходимых условий дают адекватные со статистической точки зрения результаты. Понятие большой выборки определить довольно сложно и для разных тестов пороговые значения могут быть различными. Например, для теста Миза – Рогоффа Диболд и Мариано определено, что достаточный размер выборки достигается при числе прогнозных точек больше 64.

Нарушение различных предположений тестов ведет к различным потерям, и мы не будем останавливаться на этом подробно. При наличии малого числа наблюдений лучше использовать тесты знаков, поскольку в этой ситуации они дают более адекватные результаты по сравнению с другими рассмотренными тестами.

Приложение

Пусть y_{T+i} – фактическое значение показателя в момент $T+i$, $f_{T,i}$ – прогноз этого показателя в момент

T на i шагов вперед, $e_{T,i} = y_{T+i} - f_{T,i}$ – ошибка прогноза в момент T на i шагов вперед, h – горизонт прогнозирования.

1. $MAPE = 100\% \cdot \frac{1}{h} \sum_{i=1}^{h} \left| \frac{y_{T+i} - f_{T,i}}{y_{T+i}} \right| = 100\% \cdot \frac{1}{h} \sum_{i=1}^{h} \left| \frac{e_{T,i}}{y_{T+i}} \right|$

2. $MAE = \frac{1}{h} \sum_{i=1}^{h} \left| y_{T+i} - f_{T,i} \right| = \frac{1}{h} \sum_{i=1}^{h} \left| e_{T,i} \right|$

3. $RMSE = \sqrt{\frac{1}{h} \sum_{i=1}^{h} \left(y_{T+i} - f_{T,i} \right)^2} = \sqrt{\frac{1}{h} \sum_{i=1}^{h} e_{T,i}^2}$

4. *F*-тест. Тестовая статистика выглядит следующим образом:

$$F = \frac{e_B' e_A / h}{e_B' e_B / h} = \frac{e_B' e_A}{e_B' e_B} = \frac{\sum_{i=1}^{h} e_{AT,i} e_{BT,i}}{\sum_{i=1}^{h} e_{BT,i}^2} \sim F(h, h) ,$$

где

h – горизонт прогнозирования;

$e_A = \begin{pmatrix} e_{AT,1} \\ \vdots \\ e_{AT,h} \end{pmatrix}$ и $e_B = \begin{pmatrix} e_{BT,1} \\ \vdots \\ e_{BT,h} \end{pmatrix}$ – $h \times 1$ – векторы ошибок прогнозирования по моделям A и B, соответственно.

5. *Тест Моргана – Грейнджера – Ньюболда.* Тестовая статистика:

$$MGN = \frac{\hat{\rho}_{xz}}{\sqrt{\frac{1 - \hat{\rho}_{xz}^2}{h-1}}} \sim t(h-1) ,$$

где $\hat{\rho}_{xz}$ – выборочный коэффициент корреляции между суммой (x) и разностью (z) ошибок прогнозирования различных моделей.

6. *Тест Миза – Рогоффа.* Тестовая статистик:

$$MR = \frac{\hat{\gamma}_{xz}}{\sqrt{\hat{\Sigma} / h}} \overset{a}{\sim} N(0,1),$$

где $\hat{\gamma}_{xz}$ – выборочный коэффициент ковариации между суммой и разностью ошибок прогнозирования моделей A и B; $\hat{\Sigma}$ – состоятельная оценка ковариационной матрицы.

7. *Тест знаков.* В предположении о симметричности распределения разности функций потерь число положительных наблюдений в выборке размера h имеет биноминальное распределение с параметрами h и ½. Тогда тестовая статистика имеет вид:

$$S_2 = \sum_{i=1}^{h} I_+(d_i)$$

где $I_+(d_i) = \begin{cases} 1, & если \quad d_i > 0, \\ 0, & если \quad d_i \le 0, \end{cases}$, а $d_i = g(e_{AT,i}) - g(e_{BT,i})$ – разность функций потерь прогнозов A и B, g (y_{T+i}, $f_{kT,i}$) – функция потерь, характеризующая отклонения прогнозных значений показателя y_t в момент T на i шагов вперед, оцененных на основе модели k (например, по модели A, либо B), от истинного значения y_{T+i} в этот момент времени. В случае больших выборок используется статистика:

$$S_{2a} = \frac{S_2 - 0,5h}{0,5\sqrt{h}} \overset{a}{\sim} N(0,1) \ .$$

В случае отсутствия значимых различий прогнозных свойств моделей, статистика S_2 должна быть приблизительно равна $0,5h$, а S_{2a} тогда принимает значение около нуля.

8. *Ранговый тест знаков Вилкоксона.* Можно использовать, если выполняются условия симметричности разности функций потерь ошибок прогнозирования различных моделей и разность функций потерь ошибок прогнозирования является независимой одинаково распределенной случайной величиной. В этом случае тестовая статистика может быть рассчитана как:

$$S_3 = \sum_{i=1}^{h} I_+(d_i) rank|d_i|$$

где rank $|d_i|$– ранг абсолютной величины значения разности функций потерь ошибок прогнозирования различных моделей в момент времени $i = 1, ..., h$. Тогда S_3 – сумма рангов положительных значений разности функций потерь ошибок прогнозирования разных моделей. Критические значения для небольших выборок (h – мало) можно найти в специальных таблицах, для больших выборок (асимптотически) статистика имеет стандартное нормальное распределение.

Литература

1. Diebold F.X. Elements of Forecasting. – 4th ed. – Thomson South-Western, 2007.
2. Diebold F.X., Mariano R.S. Comparing Predictive Accuracy // Journal of Business and Economic Statistics. – 1995. – № 13 (3). – pp. 253–263.
3. Granger, C.W.J., Newbold P. // Forecasting Economic Time Series, Orlando, Florida: Academic Press, 1997.
4. Meese, R.A., Rogoff K. Was it Real? The Exchange Rate – Interest Differential Relation Over the Modern Floating–Rate Period // Journal of Finance. – 1997. – 43. – pp. 933–948.
5. Morgan, W.A. A test for the Significance of the Difference Between the two variances in a Sample From Normal Bivariate Population // Biometrika. – 1939–1940. – 31. – pp. 13–19.
6. Stock, J.H. and M.W. Watson (1998a) A Comparison of Linear and Non-Linear Univariate Models for Forecasting Macroeconomic Time Series, NBER WP #6607, June.
7. Wilcoxon, F. Individual Comparisons by Ranking Methods // Biometrics Bulletin. – 1945. – 1 (6). – pp. 80–83.
8. Турунцева М., Юдин А., Дробышевский С., Кадочников П., Трунин П., Пономаренко С. Некоторые подходы к прогнозированию экономических показателей. – М.: ИЭПП, 2005.
9. Турунцева М., Киблицкая Т. Качественные свойства различных подходов к прогнозированию социально-экономических показателей РФ. – М.: ИЭПП, 2010.

pn

Marina Yu. Turuntseva
Senior Researcher, Institute of Economic Policy named after E.T. Gaidar, Executive Head of Laboratory for Short-Term Forecasting

Assessment of Forecast Quality: the Simplest Methods

Abstract

The article is devoted to the quality analysis of economic performance forecasting. The author considers the simplest statistical techniques which are used to compare several predictions of a given indicator, as well as a set of tests that allow answering the question about the statistical significance of differences between these forecasts.

Keywords: forecasting, assessment of forecast quality, economic activity, economic indicators