Инновационная модель регрессионного прогноза

Моисеев Никита Александрович,
аспирант кафедры математических методов в экономике РЭУ им. Г.В. Плеханова

Ахмадеев Булат Анасович,
аспирант кафедры информационных технологий РЭУ им. Г.В. Плеханова

В статье предлагается способ, позволяющий сократить среднюю ошибку прогноза в регрессионных моделях. Основная идея метода состоит в использовании взвешенной суммы нескольких регрессионных уравнений, удовлетворяющих предпосылкам МНК и имеющих независимые остатки, вместо одного. Показано, что если все требования метода выполнены, то возможно сократить ошибку прогноза почти в два раза, используя всего три уравнения. Данный способ позволяет создавать уравнения, которые содержат больше предикторов, чем число наблюдений. Более того, метод является более эффективным во времени, чем любое использованное в нем уравнение по отдельности. Также проиллюстрировано, что метод выглядит лучше, нежели регрессия, вычисленная по тем же независимым переменным, и, таким образом, дает более точные оценки коэффициентам регрессии.
Ключевые слова: регрессионная модель, ошибки прогноза, оценка коэффициентов

Введение

Главная проблема, с которой сталкивается исследователь при моделировании социально-экономических процессов, это неопределенность относительно такой же эффективной работы созданной модели в будущем. Эта проблема частично освещена в работах James Stock и Mark Watson (2007, 2010). Другими словами, если линейная модель полностью удовлетворяет предпосылкам метода наименьших квадратов (МНК) и имеет некоторую ошибку прогнозирования, то нельзя гарантировать, что во время использования данной модели в реальном времени эта самая ошибка будет находиться хотя бы в некоторой приемлемой области. Иногда ошибка прогноза может в несколько раз превышать первоначальную, и в этом случае от такой модели весьма мало пользы. Данная ситуация имеет место в силу ряда причин. Либо неучтенные факторы изменили свои значения таким образом, что оценки коэффициентов в модели стали смещенными, либо учтенные факторы изменили степень своего влияния на выходную переменную. Также может иметь место сочетание вышеупомянутых процессов. С целью уменьшения ошибки прогнозирования исследователь может разработать модель, которая учитывает структурные сдвиги в исследуемых процессах и изменчивость коэффициентов в регрессионном уравнении, см. например Jan J.J. Groen, Richard Paap и Francesco Ravazzolo (2009) и J.H. Wright (2009). Однако следующая проблема все еще остается не до конца решенной. Во время спецификации уравнения регрессии, пытаясь сделать уравнение максимально удовлетворяющим предпосылкам МНК, исследователь может не принять во внимание большой объем данных, на самом деле оказывающих влияние на зависимую переменную модели. Вследствие вышесказанного, особо остро стоит вопрос разработки методологии спецификации регрессионных уравнений, которая могла бы охватить больше объясняющих переменных и таким образом существенно снизить ошибку прогноза, и в то же время не нарушала классических постулатов построения линий регрессии.

Эмпирические предпосылки разрабатываемой методологии

Представим, что разрабатывается модель для прогнозирования инфляции экономики США. Выберем в качестве выходной переменной квартальный индекс потребительских цен (ИПЦ). В качестве возможных зависимых переменных будем тестировать по три лага зависимой переменной y_t и каждого квартального индекса следующих макроэкономических индикаторов: ВВП, ВНД, ставка по централизованным кредитным средствам, число занятых в несельскохозяйственном секторе, цена на нефть марки Brent, промышленный индекс Dow Jones, РВВП, полный экспорт, полный импорт, скорость денежного обращения, безработица. Таким образом, уравнение регрессии может быть записано следующим образом:

$$y_{t+1} = a_0 + \sum_{j=1}^{3} a_j y_{t-j} + \sum_{i=1}^{n} \sum_{j=1}^{3} b_{ij} x_{ij} \quad (1)$$

где n - число объясняющих переменных, a_j и b_{ij} обозначают коэффициенты лагового ИПЦ и лаговых объясняющих переменных соответственно, a_0 - константа модели

В данном случае мы не принимаем во внимание нулевые лаги, так как все макроэкономические данные недоступны в самом начале следующего отчетного периода, а выпускаются только во время квартала. Поэтому представляется невозможным осуществление прогноза в самом начале квартала в случае, если разработанная модель опирается на эти данные. Для построения модели используется 95 наблюдений, начиная с первого квартала 1960 года. После того как первое уравнение рассчитано окно данных передвигается вперед на одно наблюдение и уравнение пересчитывается снова. Данная процедура повторяется 70 раз. Для спецификации уравнения регрессии решалась следующая оптимизационная задача:

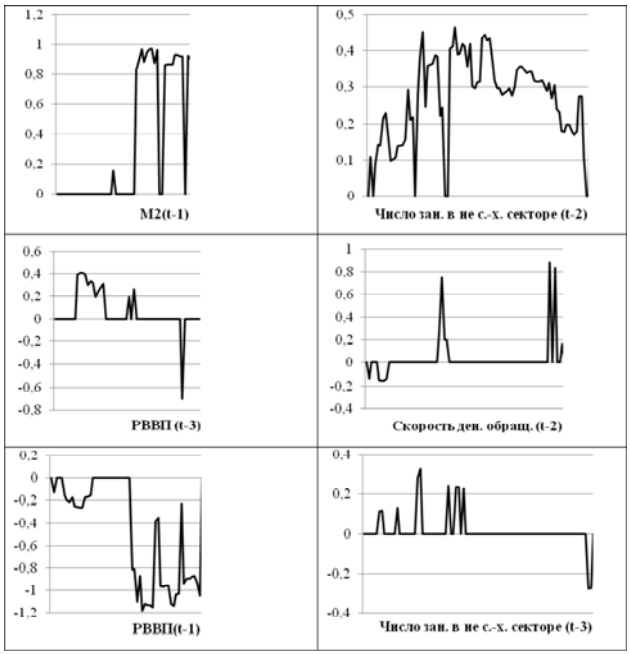
$$MSE \to \min$$
$$\begin{cases} a_k < 0{,}05; \\ m \le 10; \\ VIF_k \le 7; \\ CN \le 6; \\ 1{,}5 < DW < 2{,}5 \end{cases} \quad (2)$$

где MSE - средняя ошибка прогноза, a_k - уровень значимости для *k*-ого предиктора, $k \in [1..m]$, *m* – количество отобранных предикторов, VIF_k - фактор расширения инфляции для *k*-ого предиктора, CN – индекс обусловленности

На самом деле здесь не так важно, какой именно способ отбора переменных используется, главное, что отбираются наиболее качественное уравнение из всех возможных согласно некоторому алгоритму. Для того чтобы проиллюстрировать неустойчивость структуры регрессионного уравнения, рассмотрим динамику коэффициентов на протяжении семидесяти полученных уравнений. Результаты приведены в табл. 1.

Из таблицы видно, что существуют как более или менее устойчивые предикторы, такие как M2(t-1), число занятых в несельскохозяйственном секторе (t-2) и РВВП(t-1), так и совершенно неустойчивые (РВВП(t-3), скорость денежного обращения (t-2) и число занятых в несельскохозяйственном секторе (t-3)). Проведенный эксперимент выявил большое количество структурных сдвигов во время вычисления уравнений. Почти при каждом сдвиге окна данных могла быть построена другая регрессия, которая бы являлась лучше предыдущей. Более того, некоторые коэффициенты даже меняют свой знак. Вследствие этого можно предположить, что при опущении предикторов, на самом деле влияющих на зависимую переменную, факторы, учтенные в уравнении, принимают на себя часть их влияния на выходную переменную модели. К примеру M2(t-1) часто ведет себя как весьма значительный предиктор (коэффициент относительно высок), и тот факт что показатель M2(t-1) не учтен в уравнении с целью устранения мультиколлинеарности не означает, что данный показатель прекратил свое влияние на ИПЦ. Это влияние просто перераспределяется между коэффициентами предикторов, вошедших в модель. И если во время прогнозирования M2(t-1) начнет показывать более высокую волатильность, рассчитанная модель более не будет устойчивой, так как не включает в себя M2(t-1) в качестве объясняющей переменной. Именно поэтому исследователь должен стремиться включить в модель как можно больше зависимых переменных. Но в то же время чем больше предикторов берется для построения регрессии, тем выше риск мультиколлинеарности и больше ошибки оценок коэффициентов. Таким образом, нахождение необходимого баланса представляется несомненно важным.

Таблица 1



Проведенный эксперимент также показал, что для одного и того же окна данных можно рассчитать несколько уравнений, удовлетворяющих системе неравенств (2). Поэтому задача выбора одного из них сводится просто к применению уравнения с наименьшем показателем MSE. Однако использование последнего для построения прогноза не гарантирует наилучшее предсказание по сравнению с остальными возможными уравнениями, которые были отброшены, особенно, если показатели MSE расходятся незначительно. Например, давайте рассмотрим три возможных уравнения R_1, R_2 и R_3 для самого первого окна данных. Общая информация о каждом из них приведена в табл. 2.

Отметим, что все три уравнения удовлетворяют системе неравенств 2, и согласно выбранному алгоритму должен быть выбрано уравнение R_1 для осуществления прогноза. Однако применение R_1 для расчета будущих значений зависимой переменной далеко не всегда возвращает меньшую ошибку прогноза нежели R_2 и R_3. Для иллюстрации вышесказанного было выбрано окно данных в 30 наблюдений, начиная с первой прогнозной величины, и рассчитан показатель MSE для всех трех рассматриваемых уравнений. Данная процедура проводилась 40 раз при каждом из которых, окно данных сме-

щалось на единицу вперед. Таким образом, можно проследить, как вычисленные модели работали бы, будь они рассчитаны в 1984 году. Результаты приведены на рис. 1.

Таблица 2

Показатель	R_1	R_2	R_3
m	8	10	7
R	0.943	0.947	0.922
F-стат. знач.	0.000	0.000	0.000
DW	1.531	1.763	1.757
α	ИПЦ (t-2) – 0.001 ИПЦ (t-3) – 0.000 ЧЗНС (t-2) – 0.000 Brent (t-1) – 0.021 Импорт (t-2) – 0.020 Экспорт (t-1) – 0.000 Экспорт (t-2) – 0.021 Безр-ца (t-1) – 0.010	NFPR (t-1) – 0.000 Brent (t-1) – 0.005 M2 (t-1) – 0.000 Dow Jones (t-1) – 0.004 РВВП (t-1) – 0.000 Импорт (t-2) – 0.037 Экспорт (t-2) – 0.005 Ск. Ден. Об. (t-1) – 0.000 Безр-ца (t-2) – 0.000 Безр-ца (t-3) – 0.000	ИПЦ (t-1) – 0.000 ВВП (t-2) – 0.000 ЧЗНС (t-2) – 0.000 Brent (t-2) – 0.021 Dow Jones (t-1) – 0.003 РВВП (t-1) – 0.022 Экспорт (t-1) – 0.012
VIF	ИПЦ (t-2) – 6.882 ИПЦ (t-3) – 5.653 ЧЗНС (t-2) – 2.054 Brent (t-1) – 1.183 Импорт (t-2) – 2.882 Экспорт (t-1) – 2.303 Экспорт (t-2) – 2.377 Безр-ца (t-1) – 2.638	ЧЗНС (t-1) – 2.993 Brent (t-1) – 1.170 M2 (t-1) – 4.242 Dow Jones (t-1) – 1.679 РВВП (t-1) – 5.331 Импорт (t-2) – 3.228 Экспорт (t-2) – 2.731 Ск. Ден. Об. (t-1) – 3.003 Безр-ца (t-2) – 2.437 Безр-ца (t-3) – 1.527	ИПЦ (t-1) – 3.179 ВВП (t-2) – 4.373 ЧЗНС (t-2) – 1.753 Brent (t-2) – 1.142 Dow Jones (t-1) – 1.179 РВВП (t-1) – 1.817 Экспорт (t-1) – 1.723
CN	5.887	5.471	4.225
MSE	0.53083	0.53548	0.63767

Рис. 1

Из графика можно видеть, что MSE для R_1, R_2 и R_3 не сохраняют тот же порядок, который представлен в таблице 2 и переплетаются с течением времени. Это подчеркивает главную идею этого раздела, что если качество нескольких регрессий более или менее эквивалентно, то тогда присутствует высокий уровень неопределенности относительно выбора уравнения для осуществления прогноза.

Методология

Пусть имеется целевая переменная y и относительно большой набор независимых переменных $(x_1, x_2 \dots x_n)$. Также положим, что существует возможность вычислить *l* регрессионных уравнений $(R_1, R_2 \dots R_l)$ внутри одного окна данных и каждая R_i имеет $e_i \sim N(0;\sigma_i)$ и удовлетворяет предпосылкам МНК со значимыми предикторами. Так можно создать новое регрессионное уравнение R, используя сумму уже вычисленных.

$$\bar{\hat{y}} = \sum_{i=1}^{l} \hat{y}_i f_i \Leftrightarrow \bar{e} = \sum_{i=1}^{l} e_i f_i \qquad (3)$$

где $\hat{y}_i$ – выходная переменная R_i

Взвешивающая функция f_i берется пропорционально MSE *i*-ой регрессии.

$$f_i = \frac{s - s_i}{(l-1) \times s} \qquad (4)$$

где s_i - это несмещенная оценка σ_i и $s = \sum s_i$

Таким образом, если используются только два уравнения регрессии, то $f_1 = 1 - \frac{s_1}{s} = \frac{s_2}{s}$ и $f_2 = \frac{s_1}{s}$. Это значит, большие веса присваиваются уравнениям с меньшим значениям MSE и наоборот. Здесь предполагается, что MSE вычисленных уравнений остаются в том же порядке в процессе прогнозирования, как и в процессе спецификации модели. Переходя к показателю $\bar{e}$, который обозначает MSE для R, он имеет математическое ожидание равное 0 и среднее квадратическое отклонение $SD(\bar{e})$. Согласно (2.3) $SD(\bar{e})$ может быть представлен в следующем виде.

$$SD(\bar{e}) = \sqrt{D\left(\sum_{i=1}^{l} e_i f_i\right)} \qquad (5)$$

Обозначим коэффициент $k_i = \frac{s_i}{\min(s_i)}$. Тогда для $\min(s_i)$ $k_i = 1$, и $k_i > 1$ для других s_i. Таким образом, любое s_i может быть выражено умножением k_i на $\min(s_i)$. Поскольку целью разрабатываемого метода является формирование уравнения, которое являлось бы лучшим среди всех выведенных, введем коэффициент K, который показывает долю $SD(\bar{e})$ в $\min(s_i)$. Теперь задача сводится к минимизации K. Приемлемым интервалом для данного коэффициента является отрезок (0..1). Если $K > 1$, то тогда лучшим будет являться уже рассчитанное уравнение с минимальной величиной MSE.

$$K = \frac{SD(\bar{e})}{\min(s_i)} < 1 \quad (6)$$

Для расчета $SD(\bar{e})$ используется общая формула дисперсии суммы *l* переменных.

$$D\left(\sum_{i=1}^{l} x_i\right) = \sum_{i=1}^{l} D(x_i) + 2\sum_{1\le i<j\le l} r(x_i x_j)\sigma_i\sigma_j \quad (7)$$

Проведем следующие замены в формуле 7: $x_i = e_i f_i$ и $s_i = k_i \min(s_i)$. Тогда коэффициент K может быть представлен как в формуле ниже.

$$K = \sqrt{\sum_{i=1}^{l} k_i^2 f_i^2 + 2\sum_{1\le i<j\le l} r(e_i e_j) k_i f_i k_j f_j} \to \min \quad (8)$$

Если же разница между s_i не является существенной и можно с достаточно высокой степенью уверенности утверждать что $s_1 = s_2 = \cdots = s_l$, тогда $k_i = 1$, $f_i = \frac{1}{l}$ и (8) может быть переписано следующим образом.

$$K \approx \sqrt{\frac{1}{l}\left(1 + \frac{2}{l}\sum_{1\le i<j\le l} r(e_i e_j)\right)} \quad (9)$$

Это показывает, что для минимизации K следует отбирать уравнения либо с независимыми, либо с имеющими отрицательную зависимость остатками. Поскольку последний случай является крайне маловероятным, то будем рассматривать нулевую корреляцию ошибок прогноза, что трансформирует (9) в (10), что подчеркивает тот факт, что точность прогноза устойчиво увеличивается с ростом числа включенных уравнений.

$$K \approx \frac{1}{\sqrt{l}} \quad (10)$$

Более того, если все e_i в наборе являются независимыми одинаково распределенными случайными величинами, подчиняющимися нормальному закону распределения $N(0;\sigma_i)$, тогда $\bar{e} \sim N(0;\sigma_{\bar{e}})$.

Опишем алгоритм вывода финальной регрессии R.

1. Рассчитываются все возможные уравнения регрессии, удовлетворяющие предпосылки МНК со значимыми предикторами;
2. Выбирается уравнение с наименьшей ошибкой прогноза, и вычисляются все k_i для остальных уравнений;
3. Используется пошаговый алгоритм для выбора набора регрессионных уравнений, который минимизировал бы коэффициент K;
4. Рассчитывается регрессия *R* по формуле (3) с использованием отобранных на предыдущем шаге уравнений.

Идеальным случаем может считаться ситуация, при которой отобранные уравнения имеют равные показатели MSE и независимые остатки. Согласно (10) всего три уравнения, которые удовлетворяют данным требованиям, позволяют снизить MSE лучшего из уравнений на 42,3%. Особо отметим тот факт, что согласно представленному выше алгоритму предложенный метод позволяет создавать регрессионные уравнения, которые включают больше объясняющих переменных, чем располагаемое количество наблюдений. Метод может быть также с успехом применен к нелинейным регрессионным моделям. Особо отметим возможность полной автоматизации методологии.

Аппробация разработанной методологии

В этой части используются те же самые данные, что и в первой части для тестирования предложенной методологии. В табл. 3 приведены данные парной корреляции для остатков использованных уравнений регрессии R_1, R_2 и R_3.

Таблица 3

	Res1	Res2	Res3
Res1	1	**0.670**	**0.649**
Знч. (двусторон)	-	0.000	0.000
N	95	95	95
Res2	**0.670**	1	**0.703**
Знч. (двусторон)	0.000	-	0.000
N	95	95	95
Res3	**0.649**	**0.703**	1
Знч. (двусторон)	0.000	0.000	-
N	95	95	95

Как можно видеть из представленной выше таблицы, корреляция между остатками трех линий регрессии является хоть и значимой, но не слишком сильной. Но как будет показано в дальнейшем, даже уравнения такого качества могут устойчиво улучшить точность прогноза. Главная сложность, возникающая в случае наличия корреляции остатков, это то, что ошибки регрессии *R* могут не подчиняться нормальному закону распределения. Поэтому рассмотрим в табл. 4 тест Колмогорова-Смирнова на нормальность ошибок прогноза, чтобы убедиться в возможности использования параметрических интервальных оценок остатков модели.

Таблица 4

	Res1	Res2	Res3	Res
N	95	95	95	95
MSE	0.53083	0.53548	0.64099	0.50159
Статистика Колмогорова-Смирнова	0.730	0.467	0.794	0.452
Асимптотическая знч. (двусторон)	**0.661**	**0.981**	**0.554**	**0.987**

В нашем примере, к счастью, остатки регрессии *R* (Res) показывают даже большую нормальность, нежели любая из отобранных регрессий в отдельности (Res1, Res2, Res3). К сожалению, такая ситуация может произойти далеко не всегда, поэтому исследователь должен обращать пристальное внимание на данный факт во время использования методологии. На рис. 2 сравниваются стандартные отклонения ошибок трех регрессий, которые уже были представлены на рис. 1, с ошибками регрессии *R*, рассчитанной согласно предложенной методологии.

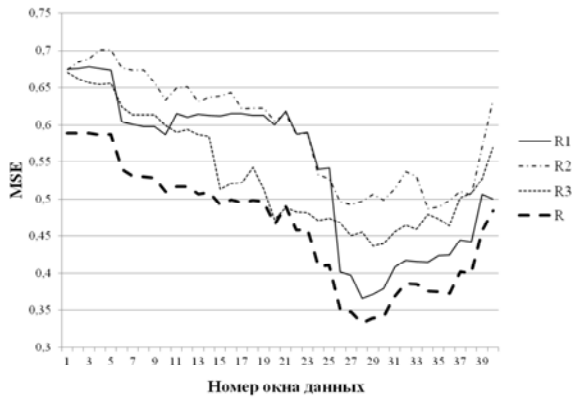


Рис. 2

Можно легко видеть, что *R* показывает систематически меньшие значения MSE на протяжении всего диапазона данных из 70 наблюдений, чем R_1, R_2 и R_3. Также обратим внимание на тот факт, что во время расчета уравнений Res был на 5,5% меньше, чем минимальная величина из Res1, Res2, Res3 (0.50159 для Res против 0.53083 для Res1), а уже во время первых 30 предсказанных значений Res стал меньше того же минимума уже на 12,2% (0.58913 для Res против 0.67063 для Res4). Регрессия *R* остается более предпочтительной, если только не происходит ради-кальных сдвигов в про-порциях MSE выбранных ре-грессий. На рис. 3.1 можно наблюдать такую ситуацию в районе 19-ого-22-ого окна данных, где Res3 располагался намного ниже, чем Res1 и Res2, что негативно сказалось на *R*.

Двигаясь дальше, для иллюстрации того факта, что метод позволяет вычислять более точные оценки коэффициентов, на рис. 3 приводится динамика MSE для регрессии R_4, построенной с помощью МНК, и которая включает те же самые предикторы, что и *R*. Также здесь показан показатель MSE для регрессии R_5, которая включает все предикторы, заявленные для возможного включения в модель в первом разделе.

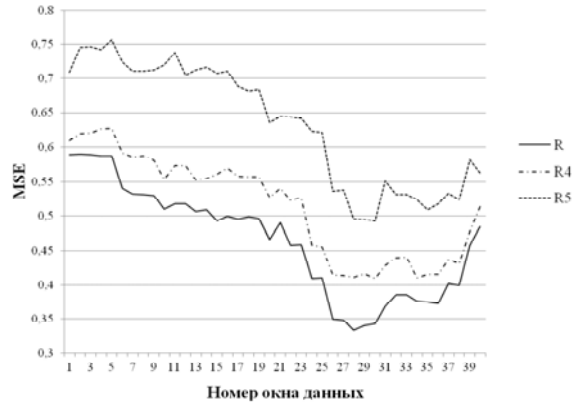


Рис. 3

Причиной того, что *R* устойчиво лучше, чем R_4 и R_5 заключается в следующем. В следствие сильной мультиколлинеарности присутствует высокий риск получения ошибочных оценок коэффициентов регрессии, которые не отражают истинные зависимости между данными. Естественно, что чем больше предикторов используется в уравнении, тем более точной получается модель. Но как только она начинает применяться для реального прогноза, она возвращает намного большие ошибки, нежели уравнение с меньшим количеством предикторов, но лучшей удовлетворенностью предпосылок МНК. Динамика MSE для R_5 только подтверждает сказанное выше.

Литература

1. Aron, Janine and John Muellbauer (2010). New Methods for Forecasting Inflation, Applied to the U.S., manuscript, Nuffield College Oxford.
2. George, E. I. and R. E. McCulloch, 1993, Variable Selection Via Gibbs Sampling, Journal of the American Statistical Association 88, 881-889.
3. Groen, J. J. J. and H. Mumtaz, 2008, Investigating the Structural Stability of the Phillips Curve Relationship, Working Paper 350, Bank of England.
4. Jan J. J. Groen, Richard Paap, Francesco Ravazzolo. Real-Time Inflation Forecasting in a Changing World. Federal Reserve Bank of New York Staff Report no. 388, August 2009, Revised May 2012.
5. Justiniano, A. and G. E. Primiceri, 2008, The Time-Varying Volatility of Macroeconomic Fluctuations, American Economic Review 98, 604-641.
6. Koop, G. and S. Potter, 2007, Estimation and forecasting in models with multiple breaks, Review of Economic Studies pp. 763-789.
7. Orphanides, A. and S. van Norden, 2005, The Reliability of Ination Forecasts Based on Output Gap Estimates in Real Time, Journal of Money, Credit and Banking 37, 583-601.
8. Stock, J.H., and M.W. Watson (2007), Why Has U.S. Inflation Become Harder to Forecast?, Journal of Money, Credit, and Banking 39, 3-34.
9. Stock, J.H., and M.W. Watson (2010), Modeling Inflation After Crisis, Manuscript for Federal Reserve Bank of Kansas City Symposium, “Macroeconomic Policy: Post-Crisis and Risks Ahead,” Jackson Hole, Wyoming, August 26-28.
10. Wright, J. H., 2009, Forecasting U.S. Ination by Bayesian Model Averaging, Journal of Forecasting 28, 131-144.