УДК 519.254
DOI: http://dx.doi.org/10.21686/2500-3925-2017-3-10-20

Н.А. Моисеев
Российский экономический университет имени Г.В. Плеханова, Москва, Россия

Вычисление истинного уровня значимости предикторов при проведении процедуры спецификации уравнения регрессии

Данная научная работа посвящена новому численному методу, вычисляющему несмещенные оценки p-значений для предикторов линейных регрессионных моделей с учетом числа потенциальных объясняющих переменных, их дисперсионно-ковариационной матрицы и степени ее неопределенности, основанной на числе рассматриваемых наблюдений. Такая поправка помогает ограничивать число ошибок 1-ого рода в научных исследованиях, значительно понижая число публикаций, декларирующих ложные зависимости в качестве истинных. Сравнительный анализ с такими существующими методами как поправка Бонферрони и поправка Шехата и Уайта явным образом демонстрирует их недостатки, особенно в случае, когда число потенциальных предикторов сравнимо с числом наблюдений. Также в процессе проведения сравнительного анализа было показано, что когда дисперсионно-ковариационная матрица набора потенциальных предикторов является диагональной, т.е. данные независимы, предложенная простая поправка является лучшим и самым легким в реализации методом для получения несмещенных корректировок традиционных p-значений. Однако, в случае присутствия сильно коррелированных данных простая поправка переоценивает истинные p-значения, что может приводить к ошибкам 2-ого рода. Также было выявлено, что исправленные p-значения зависят от числа наблюдений, числа потенциальных объясняющих переменных и выборочной дисперсионно-ковариационной матрицы. Например, если имеется только две потенциальных объясняющих переменных, конкурирующие за одну позицию в регрессионной модели, тогда, если они слабо коррелированы, исправленное p-значение будет ниже, чем в случае когда число наблюдений меньше и наоборот; если данные сильно коррелированы, случай с большим числом наблюдений будет показывать более низкое исправленное p-значение. С увеличением корреляции все поправки независимо от числа наблюдений стремятся к исходному p-значению. Данный феномен легко объяснить: с приближением коэффициента корреляции к единице две переменных практически линейно зависят друг от друга и в случае, если одна из них является значимой, то и другая почти наверняка будет демонстрировать такую же значимость. С другой стороны, если выборочная дисперсионно-ковариационная матрица стремится к диагональной и число наблюдений стремится к бесконечности, то предложенный численный метод будет возвращать поправки, близкие к простой поправке. В случае, когда число наблюдений много больше числа потенциальных предикторов, тогда поправка Шехата и Уайта дают примерно одинаковые поправки с предложенным численным методом. Однако, в намного более распространенных случаях, когда число наблюдений сравнимо с числом потенциальных предикторов, существующие методы демонстрируют достаточно значительные неточности. Когда число потенциальных предикторов больше доступного числа наблюдений, представляется невозможным рассчитать истинные p-значения. Вследствие этого рекомендуется не рассматривать такие наборы данных при построении регрессионных моделей, поскольку только выполнение вышеупомянутого условия обеспечивает расчет несмещенных корректировок p-значения. Предлагаемый метод полностью алгоритмизирован и может быть внедрен в любой пакет статистического анализа данных.

Ключевые слова: *регрессионные модели, корректировка p-значений, значимость предикторов, численный метод, распределение Уишарта, дисперсионно-ковариационная матрица, преобразование Холецкого.*

Nikita A. Moiseev
Plekhanov Russian University of Economics, Moscow, Russia

Calculating the true level of predictors significance when carrying out the procedure of regression equation specification

The paper is devoted to a new randomization method that yields unbiased adjustments of p-values for linear regression models predictors by incorporating the number of potential explanatory variables, their variance-covariance matrix and its uncertainty, based on the number of observations. This adjustment helps to control type I errors in scientific studies, significantly decreasing the number of publications that report false relations to be authentic ones. Comparative analysis with such existing methods as Bonferroni correction and Shehata and White adjustments explicitly shows their imperfections, especially in case when the number of observations and the number of potential explanatory variables are approximately equal. Also during the comparative analysis it was shown that when the variance-covariance matrix of a set of potential predictors is diagonal, i.e. the data are independent, the proposed simple correction is the best and easiest way to implement the method to obtain unbiased corrections of traditional p-values. However, in the case of the presence of strongly correlated data, a simple correction overestimates the true p-values, which can lead to type II errors. It was also found that the corrected p-values depend on the number of observations, the number of potential explanatory variables and the sample variance-covariance matrix. For example, if there are only two potential explanatory variables competing for one position in the regression model, then if they are weakly correlated, the corrected p-value will be lower than when the number of observations is smaller and vice versa; if the data are highly correlated, the case with a larger number of observations will show a lower corrected p-value. With increasing correlation, all corrections, regardless of the number of observations, tend to the original p-value. This phenomenon is easy to explain: as correlation coefficient tends to one, two variables almost linearly depend on each other, and in case if one of them is significant, the other will almost certainly show the same significance. On the other hand, if the sample variance-covariance matrix tends to be diagonal and the number of observations tends to infinity, the proposed numerical method will return corrections close to the simple correction. In the case when the number of observations is much greater than the number of potential predictors, then the Shehata and White corrections give approximately the same corrections with the proposed numerical method. However, in much more common cases, when the number of observations is comparable to

the number of potential predictors, the existing methods demonstrate significant inaccuracies. When the number of potential predictors is greater than the available number of observations, it seems impossible to calculate the true p-values. Therefore, it is recommended not to consider such datasets when constructing regression models, since only the fulfillment of the above condition ensures calculation of unbiased p-value corrections. The proposed method is easy to program and can be integrated into any statistical software package.

Keywords: *regression models, p-value adjustment, significance of predictors, randomization method, Wishart distribution, variance-covariance matrix, Cholesky decomposition.*

Введение

Существует множество методов отбора переменных при построении множественной регрессии, начиная с традиционных подходов прямого отбора (см. [5] и [8]) и заканчивая разнообразными информационными критериями и методами взвешивания регрессионных уравнений, например [1], [2], [3], [4], [15], [16], [17]. Применимость того или иного метода в определенных ситуациях является широко обсуждаемой темой в эконометрической литературе. Наиболее заметной общей чертой методов отбора переменных в уравнение является попытка найти баланс между простотой модели и величиной наблюдаемых абсолютных отклонений. Обобщая вышесказанное, можно заключить, что мы накладываем определенный штраф на величину наблюдаемых среднеквадратических отклонений, главным образом, в зависимости от числа наблюдений и числа предикторов, включаемых в модель. Чем больше число наблюдений в сравнении с количеством включенных в модель параметров, тем меньший штраф мы накладываем на наблюдаемые среднеквадратические отклонения.

В настоящее время со стремительным развитием компьютерных технологий и систем сбора статистической информации высоким спросом пользуются системы автоматической спецификации регрессионных уравнений. Множество исследователей используют запрограммированные алгоритмы построения моделей в качестве инструмента «data-mining», тестируя огромные массивы данных, содержащие фактически каждую переменную, которая имеет хотя бы призрачные шансы влиять на рассматриваемый процесс, см. например [10], [13]. Распознав возможность возникновения ошибок 1-ого рода в результате выполнения таких алгоритмов, авторы работы [9] предложили делать выводы относительно включения той или иной переменной в модель исходя из общего числа рассматриваемых потенциальных объясняющих переменных, а не исходя из количества уже отобранных факторов. Согласно работе [6] такие широко используемые методы спецификации регрессионных уравнений, как прямой отбор, обратное исключение, лучшие подмножества и др. склонны к построению моделей с ложными взаимосвязями, включая в уравнение полностью случайные факторы, на самом деле не оказывающие никакого влияния на целевую переменную. Таким образом, можно заключить, что обильный набор статистических данных неизбежно ведет к повышенному риску неверной спецификации модели в случае применения традиционных способов спецификации уравнения.

Как было явно показано в работе [7], использование методов отбора переменных ведет к получению случайного числа объясняющих переменных в конечной модели. Более того, если мы оцениваем специфицированную модель, то мы делаем вывод исходя из предположения, что отобранные факторы являются «сырыми» данными и изначально были заданы исследователем. В работе [11] утверждается, что при таких условиях традиционные тесты могут с высокой долей вероятности давать неверный результат. В контексте пошагового отбора переменных для регрессионного уравнения в работе [6] говорится дословно следующее: «when many tests of significance are computed in a given experiment, the probability of making at least one Type I error in the set of tests, that is, the maximum familywise Type I error rate (MFWER), is far in excess of the probability associated with any one of the tests» (с. 269), что означает, что в случае проведения множественных тестов на уровень значимости, вероятность совершить хотя бы одну ошибку 1-ого рода превышает аналогичную вероятность по каждому из проводимых тестов по отдельности.

Говоря о значимости каждого конкретного предиктора в линейных регрессионных моделях, сравнительно малый объем литературы посвящен проведению корректировок p-значения в зависимости от числа и ковариационной структуры вероятных объясняющих переменных. В большинстве статистических пакетов p-значение вычисляется согласно простой процедуре с использованием t-распределения независимо от размера и характеристик заданного набора данных. Учитывая сказанное выше, такой способ определения значимости предикторов является существенным опущением значимой информации, поскольку вычисление неверного уровня значимости ведет к повышенной вероятности совершения ошибки 1-ого рода при спецификации регрессионного уравнения. Как следствие, неверно построенная модель может привести исследователя к не-

правильным выводам при интерпретации ее параметров и, таким образом, оказаться убыточной для применяющих ее субъектов. Несмотря на то, что некоторые исследователи прибегают к поправкам Бонферрони или разнообразным численным методам, см. например [14], [19], [20], эти подходы дают смещенные корректировки, что может быть критично для построения адекватной модели. Для решения данной проблемы эта научная работа посвящена разработке нового численного метода для корректировки *p*-значения, который будет выдавать несмещенные поправки и, следовательно, поможет исследователям избежать совершения излишних ошибок 1-ого и 2-ого рода при построении регрессионной модели.

В данной работе рассматриваются наиболее широко распространенные методы вычисления *p*-значения, разрабатывается авторский численный метод для вычисления истинного *p*-значения для каждого предиктора модели с учетом характеристик заданного набора статистических данных и проводится имитационное тестирование разработанного метода и сравнение его результатов с существующими подходами.

Обзор методов вычисления *p*-значения

Положим, что $\{y_t, X_t : t = 1, ..., n\}$ является рассматриваемой выборкой действительных чисел, где y_t – целевая переменная, а $X_t = (1, x_{1t}, x_{2t}, ...)$ – конечный вектор потенциальных объясняющих переменных. Также предположим, что можно специфицировать простую линейную регрессионную модель, выбрав подмножество объясняющих переменных $\widetilde{X}_t$ из изначально заданного набора факторов X_t:

$$Y = \widetilde{X}B + e, \qquad (1)$$

где

$$\widetilde{X} = \begin{pmatrix} \widetilde{X}_n \\ \widetilde{X}_{n-1} \\ \vdots \\ \widetilde{X}_1 \end{pmatrix}, \quad Y = \begin{pmatrix} y_n \\ y_{n-1} \\ \vdots \\ y_1 \end{pmatrix}$$

и вектор параметров может быть вычислен в явном виде как показано ниже:

$$B = \left(\widetilde{X}^T \widetilde{X}\right)^{-1} \widetilde{X}^T Y. \qquad (2)$$

Здесь, естественно, будем предполагать, что выполняются 5 предпосылок метода наименьших квадратов (МНК).

Предпосылка 1: Строгая экзогенность ошибок, т.е. $E(\varepsilon_t|X) = 0$. Это значит, что ошибки модели не зависят от объясняющих переменных;

Предпосылка 2: Гомоскедастичность ошибок, т.е. $E(\varepsilon_t^2|X) = \sigma^2$. Дисперсия случайных отклонений является константой и не зависит от величины значений объясняющих переменных. Отметим, что невыполнение этой предпосылки называется гетероскедастичностью;

Предпосылка 3: Нормальность ошибок, т.е. $\varepsilon_t \sim N(0; \sigma)$. Случайные отклонения истинных значений зависимой переменной от модельных подчиняются нормальному распределению с нулевым математическим ожиданием и некоторой дисперсией;

Предпосылка 4: Отсутствие полной мультиколлинеарности, т.е. X^TX является положительно определенной матрицей. Здесь имеется ввиду, что среди объясняющих переменных нет функциональной линейной связи;

Предпосылка 5: Отсутствие автокорреляции остатков, т.е. $\mathrm{cov}(\varepsilon_i; \varepsilon_j) = 0$, $\forall i \neq j$. Случайные отклонения являются полностью независимыми друг от друга, что означает отсутствие систематической взаимосвязи между любыми отдельно взятыми ошибками модели.

В случае, если указанные предпосылки МНК выполняются, то общепринятой практикой является вычисление двухсторонней значимости с использованием квантилей *t*-распределения применительно к отношению величины соответствующего коэффициента к его среднеквадратическому отклонению.

$$p_i = 2 \cdot \left\{1 - T_{n-m-1}\left(\frac{|b_i|}{\sqrt{Var(b_i)}}\right)\right\}. \qquad (3)$$

В данном случае $T_{n-m-1}(x)$ – интегральная функция распределения Стьюдента с числом степеней свободы $n - m - 1$, а несмещенная оценка дисперсии коэффициентов вычисляется как:

$$Var(b_{i-1}) = s^2\left(X^T X\right)^{-1}_{ii}. \qquad (4)$$

Однако, главным вопросом в данном контексте является то, что именно олицетворяет собой *p*-значение. А обозначает оно то, что если данный предиктор будет непрерывно генерироваться и подставляться в регрессионную модуль, то мы будем наблюдать уровень надежности 95% или выше с вероятностью 5%. Таким образом, получая значимый коэффициент, обычно утверждается, что-либо случилась маловероятная ситуация, что ложный предиктор был квалифицирован как истинный, либо истинный предиктор был действительно обнаружен верно. Однако, данные умозаключения верны только тогда, когда модель не подвергалась предварительной процедуре спецификации и оценивается по сырым изначальным данным. В случае проведения процедуры отбора подмножества исходных данных для построения наилучшей модели необходимо корректировать *p*-значения коэффициентов модели исходя из числа наблюдений, числа потенциальных предикторов и их дисперсионно-ковариационной матрицы.

Именно эта идея была представлена в работе [14], в которой авторами был предложен численный метод для опреде-

ления значимости всего уравнения регрессии по наилучшему подмножеству исходного массива данных. Отметим, что предложенный метод может быть также применен к рассматриваемой задаче проведения корректировок *p*-значения. Дадим краткий обзор сути предложенной методики.

Рассмотрим набор потенциальных предикторов разрабатываемой модели $X_t = (1, x_{1t}, x_{2t}, ...)$ и обозначим p_i как *p*-значение для *i*-ого предиктора из отобранного подмножества объясняющих переменных $\tilde{X}$. Далее обозначим массив данных $\breve{X}$, который остался после проведения процедуры отбора наилучшего подмножества для модели с добавлением к нему *i*-ого предиктора из $\tilde{X}$. Таким образом, $\breve{X}$ будет включать все переменные из X, которые не были включены в $\tilde{X}$ плюс *i*-ый предиктор, находящийся под рассмотрением. После этого рассмотрим ситуацию, когда порядок наблюдений потенциальных независимых переменных из $\breve{X}$ случайным образом перемешан с фиксированием целевой переменной на своей изначальной позиции y_t, $\breve{x}_{1t}$, $\breve{x}_{2t}$, ... $\rightarrow$ y_t, $\breve{x}_{1k}$, $\breve{x}_{2k}$, ... Данное случайное преобразование наблюдений по предикторам обеспечивает точное сохранение выборочной корреляционной структуры между предикторами в перемешанном наборе данных. Также рассматриваемое преобразование обеспечивает стохастическую независимость целевой переменной y_t и перемешанного набора потенциальных независимых переменных, которые демонстрируют корреляцию только посредством случайно полученного порядка наблюдений по предикторам. Тогда процедура отбора рассматриваемого предиктора может быть осуществлена на новом перемешанном массиве данных. Обозначим q_i как *p*-значение *i*-ого предиктора из перемешанного набора данных. Если $q_i < p_i$, тогда решение, найденное по перемешанному набору демонстрирует лучшую подгонку к данным, нежели изначально заданные переменные. Для любого заданного набора данных можно произвести достаточно большое число перемешиваний из чего следует, что пропорция случаев, когда $q_i < p_i$ оценивается путем проведения имитаций. Полученная оценка будет являться исправленным *p*-значением для определения статистической значимости *i*-ого предиктора. Для заданного набора данных увеличение числа случайных перемешиваний будет отражаться в увеличении точности оцениваемого значения.

Описанная выше процедура может быть сведена к следующему алгоритму:

1. Определить исследуемый предиктор и записать соответствующее *p*-значение p_i.
2. Установить счетчик KOUNT = 0
3. DO n = 1 TO N
   а) Случайным образом перемешать $\breve{x}_{1t}$, $\breve{x}_{2t}$, ... независимо от y_t т.е. y_t, $\breve{x}_{1t}$, $\breve{x}_{2t}$, ... $\rightarrow$ y_t, $\breve{x}_{1k}$, $\breve{x}_{2k}$, ...
   б) Для перемешанного набора данных определить значение $K = 1$, если существует хотя бы один предиктор, чье *p*-значение меньше *p*-значения исследуемого предиктора, а именно $q_i < p_i$. Иначе, определить $K = 0$
   в) KOUNT = KOUNT+K
4. ENDDO
5. Скорректированное *p*-значение = KOUNT/N

В результате процесса перемешивания все возможные комбинации являются равновероятными. Следовательно, если исследуемый предиктор генерируется согласно системе, где он на самом деле не связан с целевой переменной, тогда наблюдаемое *p*-значение p_i с одинаковой долей вероятности будет таким же по величине как и любое *p*-значение q_i, полученное путем случайного перемешивания.

Еще один метод, который может рассматриваться для проведения корректировок изначального *p*-значение называется поправкой Бонферрони, которая была названа в честь итальянского математика Карло Эмилио Бонферрони. Тестирование статистических гипотез основано на отвержении нулевой гипотезы в случае, если вероятность получения наблюдаемых статистических данных при истинности нулевой гипотезы сравнительно мала. Однако, если проводятся множественные сравнения или тестируются множественные гипотезы, то шанс появления редкого события возрастает и поэтому вероятность неверно отвергнуть нулевую гипотезу (т.е. совершить ошибку 1-ого рода) возрастает, см. работу [12]. В основе поправки Бонферрони лежит идея того, что если исследователь тестирует m гипотез, тогда для фиксирования вероятности совершить ошибку 1-ого рода каждая отдельная гипотеза тестируется на уровне значимости $1/m$ помноженное на требуемый совокупный уровень значимости.

Применим данную идею к корректированию *p*-значения. Если желаемый уровень значимости предиктора равен p_i, тогда после наложения поправки Бонферрони исследуемый предиктор будет тестироваться на уровень значимости p_i/m. Например, если имеется $m = 10$ потенциальных кандидатов на место исследуемого предиктора и желаемый уровень значимости $p_i = 0{,}05$, тогда после проведения поправки Бонферрони необходимо тестировать потенциальные предикторы на уровень значимости $p_i = 0{,}05/10 = 0{,}005$.

Метод получения несмещенных корректировок *p*-значения

Для начала рассмотрим простейший случай, когда дисперсионно-ковариационная мат-

рица $\breve{X}$ является диагональной матрицей, что подразумевает отсутствие корреляции между потенциальными предикторами. Тогда скорректированное *p*-значение для рассматриваемого предиктора может быть аналитически вычислено, как показано ниже:

$$p_{adj} = 1 - (1 - p_i)^m, \qquad (5)$$

где m – число предикторов в матрице $\breve{X}$.

Формула (5) выводится от обратной вероятности. Поскольку любой предиктор из матрицы $\breve{X}$ потенциально мог бы занять место рассматриваемого предиктора, для корректировки исходного *p*-значения необходимо вычислить вероятность того, что хотя бы одно исходное *p*-значение анализируемых предикторов попадет в рассматриваемый квантиль, так как при проведении процедуры спецификации регрессионного уравнения выбирается естественно наилучший из предикторов. Чем больше потенциальных предикторов имеется в матрице $\breve{X}$, тем выше шанс найти среди них тот, который хорошо описывает целевую переменную, даже если все данные были случайным образом сгенерированы и не имеют никакой зависимости с выходной переменной y_t. Именно поэтому скорректированное *p*-значение весьма сильно отличается от исходного в случае, если число потенциальных объясняющих переменных достаточно большое. Рис. 1 в явном виде иллюстрирует на каком уровне исходного *p*-значения необходимо тестировать рассматриваемый предиктор, чтобы получить определенную значимость при наличии заданного количества потенциальных предикторов. Требуемый уровень такого исходного *p*-значения вычисляется с помощью выражения p_i из (5).

$$p_i = 1 - \sqrt[m]{1 - p_{adj}}\ . \qquad (6)$$

Из рис. 1 можно видеть, к примеру, что для обеспечения

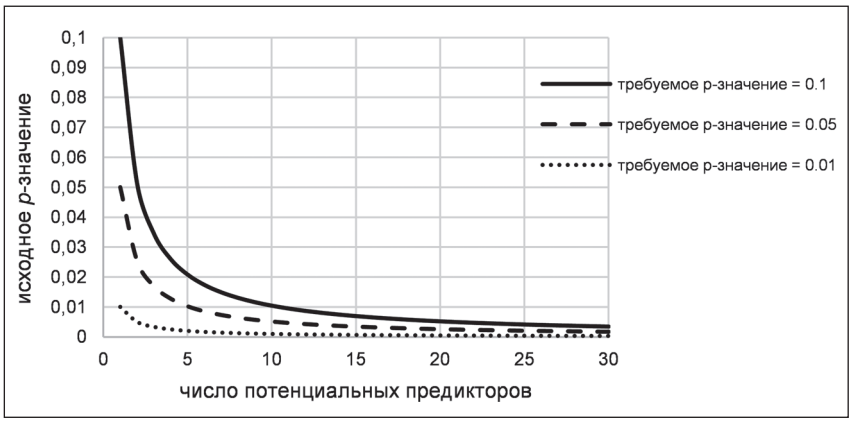

Рис. 1. Уровень исходного *p*-значения в зависимости от числа предикторов соответствующий заданной истинной значимости

истинного уровня значимости $p_{adj} = 0{,}1$, необходимо тестировать рассматриваемый предиктор на исходном уровне значимости $p_i = 0{,}007$ при числе потенциальных предикторов равному пятнадцати. В случае, если $p_{adj} = 0{,}05$, то при том же количестве предикторов требуемый уровень исходного *p*-значения должен равняться 0,0034. Таким образом, можно заключить, что в случае не проведения корректировок, вероятность совершения ошибки 1-ого рода стремительно возрастает с ростом количества потенциальных предикторов в $\breve{X}$.

Здесь также отметим, что в случае, если требуемый уровень значимости относительно низок ($p_{adj} < 0{,}1$), тогда поправка Бонферрони способна с приемлемой точностью заменить простую корректировку *p*-значения, представленную в (5), поскольку из (6) имеем $p_i \approx p_{adj}/m$. Однако, в случае, если $p_{adj} > 0{,}5$, ошибка поправки Бонферрони становится значительной и должна быть учтена соответствующим образом.

Далее рассмотрим ситуацию, когда дисперсионно-ковариационная матрица $\breve{X}$ не является диагональной, т.е. данные имеют корреляционные взаимосвязи.

$$\Sigma = \begin{vmatrix} \sigma_1^2 & \mathrm{cov}_{12} & \cdots & \mathrm{cov}_{1m} \\ \mathrm{cov}_{21} & \sigma_2^2 & \cdots & \mathrm{cov}_{12} \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{cov}_{m1} & \mathrm{cov}_{m2} & \cdots & \sigma_m^2 \end{vmatrix}$$

В данном случае предлагается прибегнуть к численной процедуре, которая базируется на машинной генерации матрицы $\breve{X}$ согласно выборочной дисперсионно-ковариационной матрицы. Здесь, дополнительно к предпосылкам 1–5, будем полагать следующее:

Предпосылка 6. Нормальность потенциальных предикторов, т.е. $x_{it} \sim N(m_i, \sigma_i)$.

Ключевым недостатком метода, предложенного в работе [14], является тот факт, что метод случайных перестановок не учитывает неопределенность, связанную с получением всего лишь несмещенной оценки истинной дисперсионно-ковариационной матрицы Σ по анализируемой выборке, что ведет к смещенности скорректированных *p*-значений. Поскольку мы имеем возможность рассчитать только выборочную дисперсионно-ковариационную матрицу, надежность предлагаемого метода зависит от пропорции числа наблюдений и числа потенциальных объясняющих переменных в $\breve{X}$. Поэтому в данном исследовании предлагается численный метод, дающий несмещенные оценки путем случайной генерации не только набора данных $\breve{X}$, но также его дисперсионно-ковариационной матрицы. Для этого, при условии выполнения предпосылки 6 можно использовать либо распределение Уишарта, либо

обратное распределение Уишарта в качестве априорного, см. [18].

$$\Sigma = W_m^{-1}\left(n-1, \breve{\Sigma}^{-1}\right)\cdot(n-1) = \\ = \frac{W_m\left(n-1, \breve{\Sigma}\right)}{n-1}, \qquad (7)$$

где n – число наблюдений,
$m \le n$ – число потенциальных объясняющих переменных в $\breve{X}$,
$\breve{\Sigma}$ – выборочная дисперсионно-ковариационная матрица для набора данных $\breve{X}$.

Таким образом, имеется возможность случайным образом сгенерировать некоторую реализацию истинной дисперсионно ковариационной матрицы при условия наличия выборочной. После проведения данной процедуры имеется возможность сгенерировать набор потенциальных объясняющих переменных $\breve{X}$ согласно некоторой полученной реализации дисперсионно-ковариационной матрицы. Для этого прибегнем к следующей процедуре. Во-первых, определим вектор-столбец независимых, идентично распределенных переменных Z, которые подчиняются нормальному распределению с нулевой средней и единичной дисперсией, что подразумевает, что

$$E\left(ZZ^T\right) = I_m.$$

После чего генерируем $\breve{X}_t$ с помощью разложения Холецкого сгенерированной дисперсионно-ковариационной матрицы Σ.

$$\breve{X}_t^T = SZ + \mu, \qquad (8)$$

где $\Sigma = SS^T$ и μ – вектор-столбец математических ожиданий соответствующих потенциальных предикторов.

Ниже приведем доказательство для формулы (8):

$$E\left\{\left(\breve{X}_t^T - \mu\right)\left(\breve{X}_t^T - \mu\right)^T\right\} = \\ = E\left(SZZ^TS^T\right) = SE\left(ZZ^T\right)S^T = \\ = SS^T = \Sigma.$$

На самом деле, для проведения имитаций не обязательно знать истинные математические ожидания μ, поскольку для расчета вектора параметров (за исключением константы модели) используются центрированные данные. Так как выборочная средняя содержит смещение и истинную среднюю, представляется рациональным генерировать $\breve{X}$, предполагая $\mu = 0$ без потери точности последующих вычислений, поскольку математические ожидания влияют только на константу модели.

$$\breve{X}_t^T = SZ. \qquad (9)$$

Сгенерировав новый набор случайных данных $\breve{X}$, подставим потенциальные предикторы один за другим на место рассматриваемой независимой переменной и присвоим единицу данному опыту в случае, если хотя бы один предиктор показал p-значение, меньшее, чем изначально полученное. По окончании одного опыта происходит повторная генерация дисперсионно-ковариационной матрицы Σ и последовательно нового набора данных $\breve{X}$. После чего повторяется процедура подставления предикторов в уравнение и опыту присваивается значение либо ноль, либо единица. Исправленное p-значение представляет собой отношение просуммированных присвоенных значений к общему числу проведенных имитаций. Таким образом, предлагаемая процедура при условии выполнения предпосылок 1–6 дает несмещенные исправленные p-значения.

Описанная выше процедура может быть алгоритмизирована следующим образом:

1. Определить исследуемый предиктор и записать соответствующее p-значение p_i.
2. Установить счетчик KOUNT = 0
3. DO n = 1 TO N
   а) Сгенерировать Σ для $\breve{X}$ согласно формуле (7)
   б) Применяя формулу (9) генерируем $\breve{X}$ согласно вновь полученной дисперсионно-ковариационной матрице Σ
   в) Для созданных фиктивных данных определим значение $K = 1$, если существует хотя бы один предиктор, чье p-значение меньше p-значения исследуемого предиктора, а именно $q_i < p_i$. Иначе, определить $K = 0$
   г) KOUNT = KOUNT+K
4. ENDDO
5. Скорректированное p-значение = KOUNT/N

Имитационный эксперимент

Проведем анализ поведения p-значений предикторов при различных формах дисперсионно-ковариационной матрицы Σ, разном числе наблюдений и числе потенциальных предикторов. Также представим сравнительный анализ предложенного метода и других широко распространенных подходов к корректировке p-значения.

Для начала исследуем случай, когда $\breve{X}$ включает в себя всего лишь две потенциальных объясняющих переменных ($m = 2$), которые имеют определенный выборочный коэффициент корреляции. На рис. 2 представлено сравнение двух кривых исправленных p-значений, рассчитанных по наборам данных разной длинны ($n = 200$, $n = 5$). В обоих случаях была произведена корректировка исходного p- значения, равного 0,05. Для получения каждой точки графика было проведено по 10 000 000 имитаций согласно предложенному алгоритму. В случае выполнения предпосылок 1–6, чем больше наблюдений имеется в рассматриваемом окне, тем более точно будет оценена истинная дисперсионно-ковариационная матрица Σ, что означает, что когда $n = 200$, мы получаем достаточно точную оценку Σ и наоборот, имеем высокий уровень неопределен-

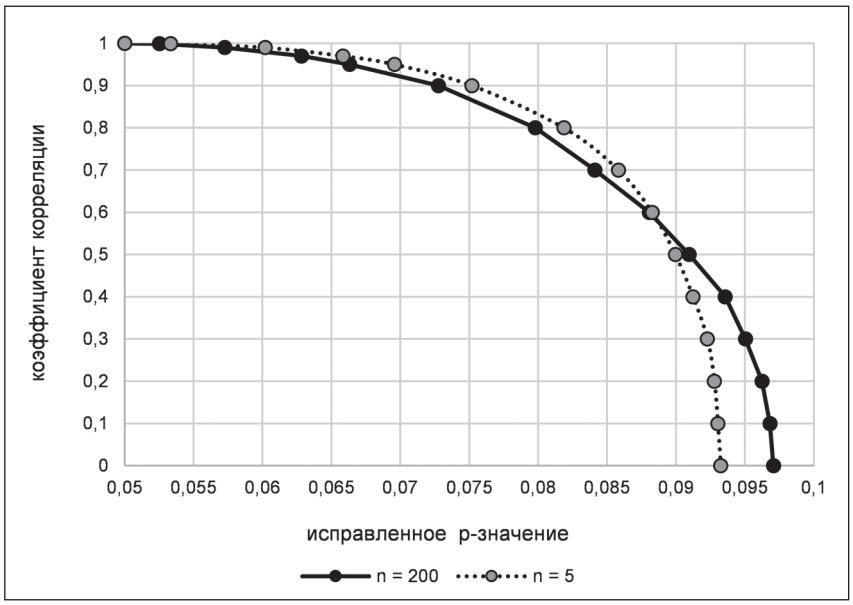

Рис. 2. Корректировка *p*-значения, равного 0,05, в случае двух потенциальных предикторов

ности относительно $\sum$ когда $n = 5$.

Поэтому можно видеть из представленного графика, что вторая кривая показывает более низкое исправленное *p*-значение, чем первая, когда коэффициент корреляции невысок и наоборот, когда присутствует значительная степень корреляции между этими двумя потенциальными предикторами. Это происходит вследствие того, что ситуация, когда при $n = 5$ выборочный коэффициент корреляции равен нулю, отнюдь не означает, что истинный коэффициент корреляции равен нулю. Имеется высокая вероятность того, что корреляция на самом деле окажется равной 0,1, 0,2 или даже 0,4 в абсолютном выражении. С другой стороны, когда $n = 200$ выборочный коэффициент корреляции является намного более точной оценкой истинного. Те же самые рассуждения можно применить к области графика, где выборочный коэффициент корреляции близок к единице. При наличии короткого окна данных не представляется возможным вычислить истинный коэффициент корреляции с достаточной точностью, и его функция плотности вероятности получается ассиметричной. Поэтому в этой области графика вторая кривая демонстрирует более высокие исправленные *p*-значения, чем первая.

Подводя итоги сказанного выше, можем заключить следующее. Первая кривая стремится к значению 0,0975 с приближением выборочного коэффициента корреляции к нулю, что является простой поправкой, представленной в формуле (5). Когда коэффициент корреляции стремится к единице, исправленные *p*- значения в обоих случаях стремятся к исходному, равному 0,05. Эти кривые пересекаются в точке, где выборочный коэффициент корреляции приблизительно равен 0,6, а его функция плотности вероятности в случае $n = 5$ меняет ассиметрию с положительной на отрицательную.

Для проведения сравнительного анализа различных подходов к вычислению исправленного *p*-значения изобразим несмещенные *p*-значения, вычисленные согласно предложенному методу (ось абсцисс) против исправленных *p*-значений, вычисленных согласно другим существующим методикам (ось ординат). В частности будем сравнивать традиционный расчет *p*-значения, поправку Бонферрони, поправку Шехата и Уайта, предложенные простую корректировку и численную поправку *p*-значения. Например, на рисунках 3*а* и 3*б* представлено сравнение рассматриваемых поправок в случае, когда $n = 5$, $m = 2$ и выборочный коэффициент корреляции принимает значения, сконцентрированные вокруг нуля. Главным выводом, который можно сделать, анализируя вышеупомянутые графики,

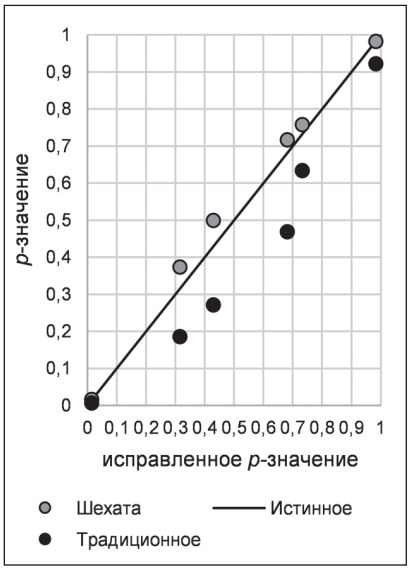

Рис. 3а. Сравнение методов корректировок *p*-значения ($n = 5$, $m = 2$)

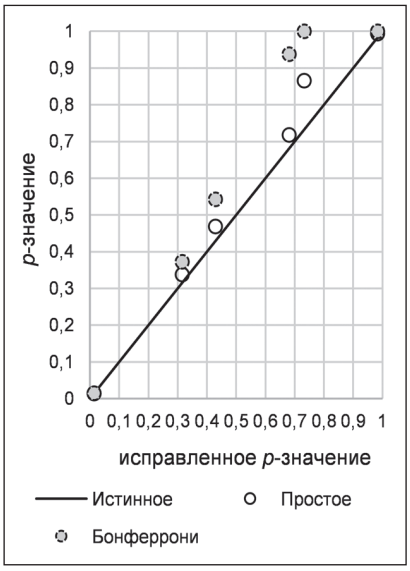

Рис. 3б. Сравнение методов корректировок *p*-значения ($n = 5$, $m = 2$)

является тот факт, что традиционный расчет *p*-значения недооценивает истинный его уровень, что выражается в повышенном риске совершения ошибки 1-ого рода, тогда как простая поправка, поправка Шехата и Уайта и поправка Бонферрони в среднем переоценивают истинные *p*-значения и, таким образом, вносят вклад в повышение риска возникновения ошибок 2-ого рода, т.е. исключения значимого предиктора из уравнения. Сравнивая поправку Шехата и Уайта с предложенным методом, можно заключить, что их оценки достаточно близки при $n \gg m$, однако, если данное требование не выполнено – поправка Шехата и Уайта дает значительно смещенные оценки *p*-значения, см. рисунки 3*а*, 3*б* и 4*а*, 4*б*.

На рисунках 4*а* и 4*б* наглядно демонстрируется тот факт, что в случае, если количество наблюдений *n* достаточно мало и число потенциальных объясняющих переменных *m* примерно равно *n*, тогда существующие методы возвращают более смещенные оценки истинного *p*-значения, чем в первом рассмотренном примере. Следует отметить, что оценки *p*-значений, полученные по методу Шехата и Уайта, дают смещение вверх при относительно слабой выборочной корреляции и смещение вниз при сильной. Такой эффект проявляется вследствие того, что метод, разработанный Шехата и Уайтом не учитывает вариативность истинного коэффициента корреляции относительно выборочного при определенном числе наблюдений.

Однако, ситуация, представленная на рисунках 4*а* и 4*б*, не является типичной для проводимых в экономике исследований. Поэтому рассмотрим более правдоподобный пример, когда в распоряжении исследователя имеются данные по $m = 25$ потенциальных предикторов, имеющим по $n = 30$

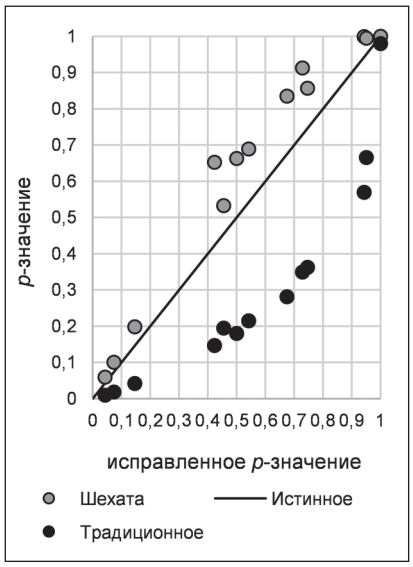

Рисунок 4а. Сравнение методов корректировок *p*-значения ($n = 6$, $m = 5$)

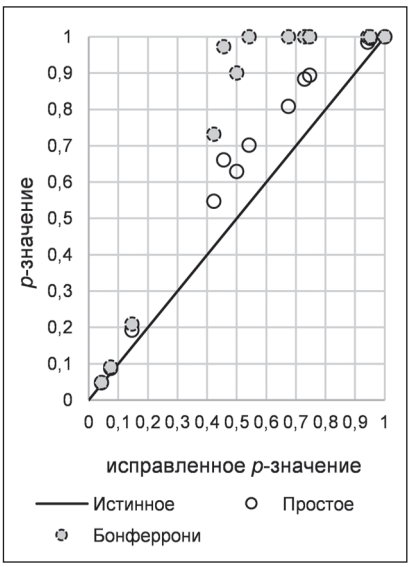

Рисунок 4б. Сравнение методов корректировок *p*-значения ($n = 6$, $m = 5$)

наблюдений. Результаты полученных расчетов по рассматриваемым методам корректировки *p*-значений в данном случае представлены на рисунках 5*а* и 5*б*.

Как можно видеть, традиционный расчет *p*-значения показывает наихудшую точность, что не удивительно, поскольку число потенциальных предикторов достаточно высоко. Однако, альтернативные методы также демонстрируют существенное смещение, что увеличивает вероятность совершения ошибок 2-ого рода в процессе спецификации регрессионного уравнения.

Здесь стоит особо отметить случай, когда $m > n$. В данной ситуации, если попытаться сгенерировать всю матрицу $\breve{X}$ согласно предложенному методы, то дисперсионно-ковариационная матрица $\sum$ будет необратима поскольку предикторы с индексами выше *m*-ого будут линейно выражаться через оставшиеся объясняющие переменные, что не является истиной для реальных эконо-

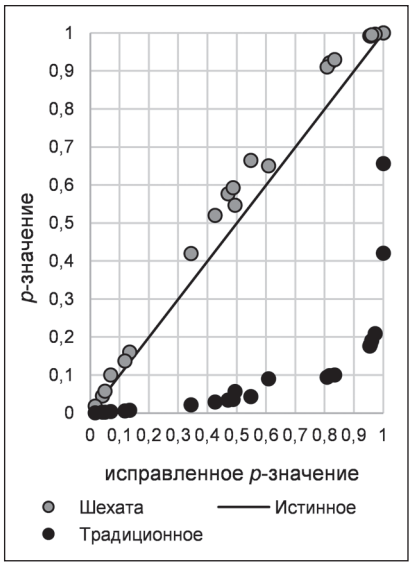

Рисунок 5а. Сравнение методов корректировок *p*-значения ($n = 30$, $m = 25$)

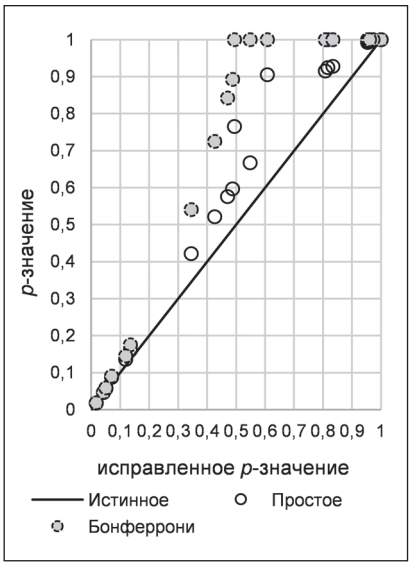

Рисунок 5б. Сравнение методов корректировок *p*-значения ($n = 30$, $m = 25$)

мических данных. Теоретически есть возможность сгенерировать матрицу $\tilde{X}$, но при этом полученные поправки не будут отражать адекватные зависимости между рассматриваемыми переменными и, таким образом, не могут считаться надежными. Простая корректировка тем не менее может применяться в случае, если есть веские основания считать, что $\sum$ является диагональной матрицей, что практически недостижимо для экономических данных. Поправка Бонферрони и корректировка Шехата и Уайта также теоретически могут применяться, но будут возвращать поправки не лучше, чем предлагаемый метод. Исходя из вышесказанного, крайне рекомендуется рассматривать набор объясняющих переменных X где $n \geq m$ для построения регрессионной модели, поскольку только выполнение данного условия обеспечивает получение несмещенных корректировок исходных p-значений.

Заключение

Представленный в данной работе численный метод возвращает несмещенные поправки исходных p-значений для объясняющих переменных, который напрямую относится к выводам относительно степени влияния этих переменных на целевую. Проведя сравнительный анализ предложенного метода и уже существующих, таких как традиционный расчет p-значения, поправка Бонферрони и поправка Шехата и Уайта, можно сделать следующие основные выводы.

1. В случае, когда дисперсионно-ковариационная матрица набора потенциальных предикторов является диагональной, т.е. данные независимы, предложенная простая поправка является лучшим и самым легким в реализации методом для получения несмещенных корректировок традиционных p-значений. Однако, в случае присутствия сильно коррелированных данных простая поправка переоценивает истинные p-значения, что может приводить к ошибкам 2-ого рода.

2. Исправленные p-значения зависят от числа наблюдений, числа потенциальных объясняющих переменных и выборочной дисперсионно-ковариационной матрицы. Например, если имеется только две потенциальных объясняющих переменных, конкурирующие за одну позицию в регрессионной модели, тогда, если они слабо коррелированы, исправленное p-значение будет ниже, чем в случае когда число наблюдений меньше и наоборот; если данные сильно коррелированы, случай с большим числом наблюдений будет показывать более низкое исправленное p-значение. С увеличением корреляции все поправки независимо от числа наблюдений стремятся к исходному p-значению. Данный феномен легко объяснить: с приближением коэффициента корреляции к единице две переменных практически линейно зависят друг от друга и в случае, если одна из них является значимой, то и другая почти наверняка будет демонстрировать такую же значимость. С другой стороны, если выборочная дисперсионно-ковариационная матрица стремится к диагональной и число наблюдений стремится к бесконечности, то предложенный численный метод будет возвращать поправки, близкие к простой поправке.

3. В случае, когда число наблюдений много больше числа потенциальных предикторов, тогда поправка Шехата и Уайта дают примерно одинаковые поправки с предложенным численным методом. Однако, в намного более распространенных случаях, когда число наблюдений сравнимо с числом потенциальных предикторов, существующие методы демонстрируют достаточно значительные неточности.

4. Когда число потенциальных предикторов больше доступного числа наблюдений, представляется невозможным рассчитать истинные p-значения. Вследствие этого рекомендуется не рассматривать такие наборы данных при построении регрессионных моделей, поскольку только выполнение вышеупомянутого условия обеспечивает расчет несмещенных корректировок p-значения.

Расчет истинных p-значений может оказать помощь в ограничении числа ошибок 1-ого рода в научных исследованиях, значительно снижая количество публикаций, декларирующих ложные зависимости в качестве истинных. Предложенные методы легко алгоритмизируются и могут быть интегрированы в абсолютное большинство существующих статистических пакетов.

Литература

1. *Akaike H.* Information theory and an extension of the maximum likelihood principle. In: Petroc B., Csake F. (Eds.) Second International Symposium on Information Theory. 1973.

2. *Akaike H.* A Bayesian extension of the minimum AIC procedure of autoregressive model fitting // Biometrika. 1979. 66. P. 237–242.

References

1. *Akaike H.* Information theory and an extension of the maximum likelihood principle. In: Petroc B., Csake F. (Eds.) Second International Symposium on Information Theory. 1973.

2. *Akaike H.* A Bayesian extension of the minimum AIC procedure of autoregressive model fitting // Biometrika. 1979. 66. P. 237–242.

3. *Bates J.M., Granger, C.W.J.* The combination of forecasts // Operations Research Quarterly. 1969. 20. P. 451–468.

4. *Buckland S.T., Burnham K.P., Augustin, N.H.* Model selection: An integral part of inference // Biometrics. 1997. 53. P. 603–618.

5. *Canning F.L.* 1959. Estimating load requirements in a job shop // Journal of Industrial Engineering. 1959. 10. P. 447.

6. *Derksen S., Keselman H.J.* Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables // British Journal of Mathematical and Statistical Psychology. 1992. 45. P. 265–282.

7. *Hurvich C.M., Tsai C.L.* The impact of model selection on inference in linear regression // The American Statistician. 1990. 44. 3. P. 214–217.

8. *Kramer C.Y.* Simplified computations for multiple regression // Industrial Quality Control. 1957. 13. 8. 8.

9. *Larzelere R.E., Mulaik S.A.* Single-sample tests for many correlations // Psychological Bulletin. 1977. 84. P. 557 – 569.

10. *Lovell M.C.* Data mining. The Review of Economics and Statistics. 1983. 65. P. 1–12.

11. *Miller A. J.* Selection of subsets of regression variables (with discussion) // Journal of the Royal Statistical Society. 1984. A. 147. P. 389–425.

12. *Mittelhammer Ron C., Judge George G., Miller Douglas J.* Econometric Foundations. Cambridge University Press. 2000. P. 73–74.

13. *Moiseev N.A.* Linear model averaging by minimizing mean-squared forecast error unbiased estimator // Model Assisted Statistics and Applications. 2016. Vol. 11, No 4, P. 325–338.

14. *Shehata Yasser A., White Paul A* Randomization Method to Control the Type I Error Rates in Best Subset Regression // Journal of Modern Applied Statistical Methods. 2008. 7. 2. P. 398–407.

15. *Shibata Ritaei.* Asymptotically efficient selection of the order of the model for estimating parameters of a linear process // Annals of Statistics. 1990. 8. Pp. 147–164.

16. *Shibata Ritaei.* An optimal selection of regression variables // Biometrika. 1981. 68. P. 45–54.

17. *Shibata Ritaei.* Asymptotic mean efficiency of a selection of regression variables // Annals of the Institute of Statistical Mathematics. 1983. 35. P. 415–423.

18. *Wishart J.* The generalized product moment distribution in samples from a normal multivariate population // Biometrica. 1928. 20A. P. 32–52.

19. *Глазьев С.* Проблемы прогнозирования макроэкономической динамики // Российский

3. *Bates J.M., Granger, C.W.J.* The combination of forecasts // Operations Research Quarterly. 1969. 20. P. 451–468.

4. *Buckland S.T., Burnham K.P., Augustin, N.H.* Model selection: An integral part of inference // Biometrics. 1997. 53. P. 603–618.

5. *Canning F.L. 1959.* Estimating load requirements in a job shop // Journal of Industrial Engineering. 1959. 10. P. 447.

6. *Derksen S., Keselman H.J.* Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables // British Journal of Mathematical and Statistical Psychology. 1992. 45. P. 265–282.

7. *Hurvich C.M., Tsai C.L.* The impact of model selection on inference in linear regression // The American Statistician. 1990. 44. 3. P. 214–217.

8. *Kramer C.Y.* Simplified computations for multiple regression // Industrial Quality Control. 1957. 13. 8. 8.

9. *Larzelere R.E., Mulaik S.A.* Single-sample tests for many correlations // Psychological Bulletin. 1977. 84. P. 557 – 569.

10. *Lovell M.C.* Data mining. The Review of Economics and Statistics. 1983. 65. P. 1–12.

11. *Miller A. J.* Selection of subsets of regression variables (with discussion) // Journal of the Royal Statistical Society. 1984. A. 147. P. 389–425.

12. *Mittelhammer Ron C., Judge George G., Miller Douglas J.* Econometric Foundations. Cambridge University Press. 2000. P. 73–74.

13. *Moiseev N.A.* Linear model averaging by minimizing mean-squared forecast error unbiased estimator // Model Assisted Statistics and Applications. 2016. Vol. 11, No. 4, P. 325–338.

14. *Shehata Yasser A., White Paul A* Randomization Method to Control the Type I Error Rates in Best Subset Regression // Journal of Modern Applied Statistical Methods. 2008. 7. 2. P. 398–407.

15. *Shibata Ritaei.* Asymptotically efficient selection of the order of the model for estimating parameters of a linear process // Annals of Statistics. 1990. 8. P. 147–164.

16. *Shibata Ritaei.* An optimal selection of regression variables // Biometrika. 1981. 68. P. 45–54.

17. *Shibata Ritaei.* Asymptotic mean efficiency of a selection of regression variables // Annals of the Institute of Statistical Mathematics. 1983. 35. P. 415–423.

18. *Wishart J.* The generalized product moment distribution in samples from a normal multivariate population // Biometrica. 1928. 20A. P. 32–52.

19. *Glaz'ev S.* Problemy prognozirovaniya makroekonomicheskoi dinamiki // Rossiiskii

экономический журнал. 2001. № 3. С. 76–85; № 4. С. 12–22.

20. *Крыштановский А.О.* Методы анализа временных рядов // Мониторинг общественного мнения: экономические и социальные перемены. 2000. № 2 (46). С. 44–51.

ekonomicheskii zhurnal. 2001. № 3. P. 76–85; № 4. P. 12–22. (in Russ.)

20. *Kryshtanovskii A.O.* Metody analiza vremennykh ryadov // Monitoring obshchestvennogo mneniya: ekonomicheskie i sotsial'nye peremeny. 2000. № 2 (46). P. 44–51. (in Russ.)

Сведения об авторе

Никита Александрович Моисеев
Кандидат экономических наук, доцент кафедры Математических методов в экономике
РЭУ им. Г.В. Плеханова,
Москва, Россия
Эл. почта: Moiseev.NA@rea.ru

Information about the author

Nikita A. Moiseev
Cand. Sci. (Economics), Associate Professor of the Department of Mathematical Methods in Economics
Plekhanov Russian University of Economics,
Moscow, Russia
E-mail: Moiseev.NA@rea.ru