УДК 330.4 Материал поступил в редакцию 03.10.17.

Н. Н. МАЛОВА, *канд. экон. наук, доцент*
Федеральное государственное бюджетное образовательное учреждение высшего образования «Российский государственный аграрный университет – Московская сельскохозяйственная академия имени К. А. Тимирязева», Российская Федерация, г. Москва

N. N. MALOVA, *Ph. D. of Economics, Associate professor*
Federal State Budgetary Educational Institution of Higher Education "Russian State Agrarian University – Moscow Agricultural Academy named after K. A. Timiryazev", Russian Federation, Moscow

ОБ ОДНОМ ПОДХОДЕ К РАСЧЕТУ СРЕДНЕЙ ОШИБКИ АППРОКСИМАЦИИ РЕГРЕССИОННЫХ МОДЕЛЕЙ

ABOUT ONE APPROACH TO THE CALCULATION OF THE AVERAGE ERROR OF APPROXIMATION OF THE REGRESSION MODELS

Аннотация. *Средняя ошибка аппроксимации – один из показателей, характеризующих качество регрессионной модели. В данной статье обсуждаются методы вычисления обобщающего показателя. Средняя ошибка аппроксимации может быть вычислена разными способами, приводящими к различным результатам. С чисто теоретических позиций, а также на основе апробации на условном и на реальном примерах уравнений регрессии показано, что единственно правильной методикой вычисления средней относительной линейной (по модулю) ошибки аппроксимации регрессионной модели является отношение средней абсолютной ошибки (по модулю) к среднему уровню моделируемого признака.*
Ключевые слова: *средняя ошибка аппроксимации, гомоскедастичность, гетероскедастичность, регрессионная модель, относительная и абсолютная ошибки, выборочная совокупность.*

Abstract. *The average error of approximation is one of the indicators characterizing the quality of a regression model. This article discusses the methods of calculation of the generalizing indicator. The average approximation error can be calculated in different ways, leading to different results. From a purely theoretical standpoint and on the basis of tests of conditional and real examples of regression equations, we have shown that the only correct method of calculating the linear average relative (absolute) error of approximation of the regression model is the ratio of the average absolute error (in absolute value) to the average level of the simulated trait.*
Keywords: *average approximation error, homoskedasticity, heteroscedasticity, regression model, relative and absolute errors, sample.*

Средняя ошибка аппроксимации – один из показателей, характеризующих качество регрессионной модели. Особенно она важна, если модель будет использоваться для прогнозирования, так как ошибка прогноза тесно связана (хотя и не идентична) со средней ошибкой аппроксимации. Ошибка аппроксимации измеряет различие между фактическими значениями результативного признака и его расчетными значениями по регрессионной модели, т. е. между Y_i и $\hat{Y}_i$ для каждой *i*-й единицы совокупности и для совокупности в целом.

В данной статье рассмотрим методы вычисления обобщающего показателя, а точнее – систему показателей, характеризующих модель: абсолютных и относительных средних ошибок аппроксимации.

Введем следующие обозначения:

1. Абсолютные ошибки аппроксимации для каждой *i*-й единицы совокупности $u_i = \hat{Y}_i - Y_i$, т. е. это разность между расчетными по модели и фактическими значениями моделируемого признака.

2. Средняя по совокупности абсолютная (линейная) ошибка аппроксимации:

$$\overline{u} = \sum_{i=1}^{n} \left| u_i \right| : n.$$

Иначе говоря, это средний модуль абсолютных индивидуальных ошибок, так как простая сумма их по определению для линейной модели и всех порядков парабол равна нулю, и для всех других типов кривых близка к нулю, если тип линии регрессии выбран правильно.

3. Средняя квадратическая ошибка аппроксимации – это средняя квадратическая величина из индивидуальных абсолютных ошибок, обычно вычисляемая с учетом степеней свободы, т. е.:

$$S_u = \sqrt{\frac{\sum_{i=1}^{n} u_i^2}{n-p}},$$

где p – число параметров линии регрессии или число факторов в многофакторной модели.

Поскольку квадратическая ошибка входит в расчет вероятности ошибки при нормальном распределении или распределении Стьюдента, а модули имеют плохие математические свойства и не позволяют перейти к вероятностным оценкам, большинство учебников рекомендуют только среднюю квадратическую ошибку.

4. Индивидуальная линейная относительная ошибка аппроксимации – это отношение абсолютной индивидуальной ошибки к значению признака данной *i*-й единицы совокупности:

$$\varepsilon = \frac{u_i}{Y_i} \text{ (часто выражается в процентах).}$$

5. Обобщающими показателями качества модели с точки зрения степени аппроксимации являются сравнимые относительные средние ошибки – линейная и квадратическая. Методика вычисления средней квадратической относительной ошибки разногласий в литературе не вызывает. Она определяется как отношение абсолютной средней квадратической ошибки к среднему уровню моделируемого признака, т. е.:

$$v_u = \frac{S_u}{\overline{Y}}.$$

Однако было бы неверно ограничиваться только средними квадратическими оценками ошибок аппроксимации. Они полезны для расчета вероятных ошибок прогнозов и при распространении результатов, полученных по выборке на генеральную совокупность.

Если исследователя интересует вопрос: на какую величину различаются в данной совокупности предприятий или иных объектов расчетные по модели значения от фактических значений моделируемого признака в их натуральном выражении, без поправок на степени свободы, и без преувеличения роли единиц с большими ошибками, возникающего при возведении ошибок в квадрат, то ответом должна и может служить средняя линейная относительная ошибка аппроксимации.

Эта ошибка может быть вычислена разными способами, приводящими к различным результатам:

А. Как средняя арифметическая величина (простая) из индивидуальных относительных ошибок, т. е.

$$\overline{\varepsilon}_А = \frac{\sum_{i=1}^{n} \varepsilon_i}{n} = \sum_{i=1}^{n} \left| \frac{Y_i' - Y_i}{Y_i} \right| : n.$$

Б. Как отношение средней абсолютной ошибки аппроксимации к среднему уровню моделируемого признака, т. е.

$$\overline{\varepsilon}_Б = \frac{\overline{u}}{\overline{Y}} = \frac{\sum_{1}^{n} \left| u_i \right| : n}{\overline{Y}} = \frac{\sum_{i=1}^{n} \left| Y_i' - Y_i \right|}{n} : \overline{Y}.$$

В чем причина различия показателей, вычисленных по способу А и по способу Б, и можно ли сделать вывод о том, какой из них следует считать правильно отражающим качество модели?

Чтобы ответить на этот вопрос, следует вспомнить, что адэкватная регрессионная модель, полученная по методу наименьших квадратов, требует, чтобы распределение ошибок аппроксимации было гомоскедастичным. Это значит, что величина ошибки должна быть независима от значений признака Y_i.

Если же ошибки зависят от величины признака, т. е. их распределение гетероскедастично, например, при возрастании значений Y_i возрастают ошибки u_i, то метод наименьших квадратов (МНК) не дает несмещенных и состоятельных оценок влияния

факторов на результативный признак, т. е. модель неэффективна, не годится.

Теперь представим себе идеальную гомоскедастичность, иначе говоря, все ошибки u_i равны между собой (таблица) и вычислим при этом условии среднюю относительную линейную ошибку аппроксимации способом А и способом Б.

Ошибки при гомоскедастичности

$\overline{Y} = 16$

№ единицы совокупности	Y_i	$\hat{Y}_i$	u_i	ε_i, %
1	5	9	+4	+80
2	6	10	+4	+67
3	10	6	-4	−40
4	14	18	+4	+29
5	15	11	−4	−27
6	20	16	−4	−20
7	28	32	+4	+14
8	30	26	−4	−13
Σ	128	128	0	90

Не происходит полного взаимопогашения относительных ошибок, что уже свидетельствует о просчете. С увеличением значений признака Y_i уменьшается значение ε_i, т. е. существует обратная зависимость относительных ошибок от величин признака – гетероскедастичность.

В реальной совокупности гомоскедастичность абсолютных ошибок не означает их полного равенства – достаточно того, чтобы в среднем абсолютные ошибки были примерно равны при низких значениях Y_i и при высоких значениях Y_i. А следовательно, все равно будет наблюдаться гетероскедастичность относительных ошибок.

Относительные ошибки будут в среднем больше при низких Y_i. Индивидуальные относительные ошибки по методу А – это доли (проценты) от заведомо неравных значений Y_i, а складывать, сравнивать и производить какие-либо иные действия над долями от разных величин недопустимо. Поэтому имеет место нарушение математических требований в способе А.

В результате, вычисленная таким способом средняя относительная ошибка аппроксимации оказывается завышенной, поскольку при ее расчете преобладание получают единицы совокупности с низкими величинами Y_i.

По данным таблицы, средняя относительная линейная ошибка равна:

$$\sum_1^8 |\varepsilon_1| : 8 = 290\,\% : 8 = 36,25\,\%.$$

Рассчитанная средняя линейная ошибка аппроксимации по способу Б:

$$\overline{\varepsilon}_Б = \frac{\sum_1^8 |u_i|}{n} : \overline{Y} = \frac{32}{8} : 16 = 0,25, \text{ или } 25\,\%.$$

При этом расчете используется только гомоскедастичная абсолютная ошибка. Она подлежит усреднению и в результате средняя абсолютная ошибка входит в расчет средней относительной ошибки. Отсутствуют нарушения требований МНК, отсутствует и сложение неоднородных величин.

Как видим, сама величина средней относительной линейной ошибки аппроксимации по способу Б меньше на 31 %, или на 11,25 процентных пункта. Квадратическая относительная ошибка по данным таблицы составляет:

$$\nu_u = S_u : \overline{Y} = \sqrt{\frac{112}{7-2}} : 16 = 4,73 : 16 = 29,5\%.$$

Тот факт, что даже квадратическая ошибка, вопреки правилу мажорантности средних, меньше линейной по способу А – лишний раз подтверждает неверность методики А.

Может показаться, что использованный в таблице пример − искусственный. Поэтому покажем, что идентичное преувеличение средней линейной ошибки аппроксимации происходит и при решении задачи регрессионного анализа по фактическим данным, а именно: при изучении зависимости валового регионального продукта (ВРП) от величины основных производственных средств в экономике, например, что не нарушает общности рассмотрения по Приволжскому федеральному округу (2013 год).

Предварительная проверка совокупности на близость к нормальному закону распределения по *X* и по *Y* дала следующие результаты: критерий Стьюдента для показателя асимметрии: $t_{as\ x}$ = 0,987; $t_{as\ y}$ = 0844. Для показателя эксцесса: $t_{ex\ x}$ = −0,593; $t_{ex\ y}$ = −1,05. Табличное значение *t* – Стьюдента для 13 степеней свободы при значимости 0,05 составляет 2,16. Все показатели много ниже критического значения, следователь-

но, гипотеза о нормальном распределении совокупности по *X* и по *Y* не отвергается. Условия применения МНК соблюдены.

Решение на ПЭВМ по программе "Statgraphics Plus 5.0" дало уравнение регрессии: $\hat{Y}_i = -13{,}12 + 3367 X_i \pm u_i$, коэффициент детерминации R^2 = 95,4 %, критерий Фишера F = 249, что в десятки раз превышает критическое значение. Программа выдает также среднее квадратическое отклонение расчетных значений $\hat{Y}_i$ от фактических Y_i, S_u, равное 9,4 млрд р. Отсюда, квадратическое относительное отклонение $\nu_u = S_u : \overline{Y} = 9{,}4 : 56{,}92 = 0{,}165$, или 16,5 %.

Линейные ошибки по модулю составляют:

абсолютная:

$$\overline{u} = \sum_1^{14} u_i : 14 = 97{,}3 : 14 = 6{,}95 \text{ млрд р.;}$$

относительная линейная ошибка аппроксимации:

по способу А:

$$\overline{\varepsilon}_A = \sum_1^{14} \varepsilon_i : 14 = 283{,}3\,\% : 14 = 20{,}33\,\%$$

по способу Б:

$$\overline{\varepsilon}_Б = \overline{u} : \overline{Y} = \frac{6{,}95}{56{,}92} = 0{,}122, \text{ или } 12{,}2\,\%.$$

Как видим, способ А резко завышает величину средней ошибки.

Покажем, что способ А нарушает и требование о гомоскедастичности отклонений. Для этого вычислим и сравним среднее значение отклонений ε_i для единиц совокупности со значением $Y_i < \overline{Y}$ и для единиц совокупности со значениями $Y_i > \overline{Y}$. Для первой группы : $\overline{\varepsilon}(I) = 26{,}04\,\%$; для второй группы: $\overline{\varepsilon}(II) = 12{,}5\,\%$. Различие этих величин существенно, что и доказывает гетероскедастичность распределения относительных ошибок ε_i.

Напротив, усредняемые по II способу абсолютные ошибки u_i для первой группы субъектов Федерации составили $|\overline{u}_1| = 5{,}8$, для второй группы – $|\overline{u}_2| = 8{,}5$; различие оказалось несущественным.

Как видим, несмотря на большую величину абсолютных ошибок в группе субъектов со значениями $Y_i > \overline{Y}$, относительные ошибки в этой группе почти вдвое меньше.

Таким образом, с чисто теоретических позиций, а также на основе апробации на условном и на реальном примерах уравнений регрессии показано, что единственно правильной методикой вычисления средней относительной линейной (по модулю) ошибки аппроксимации регрессионной модели является отношение средней абсолютной ошибки (по модулю) к среднему уровню моделируемого признака.

СПИСОК ЛИТЕРАТУРЫ

1. **Афанасьев В., Юзбашев М. М., Гуляева Т. Н.** Эконометрика. – М. : Финансы и статистика, 2005. – 256 с.

2. **Васильева Э. К., Юзбашев М. М.** Выборочный метод в социально-экономической статистике: учебное пособие (ГРИФ). – М. : Инфра-М, 2010. – 256 с.

3. **Елисеева И. И., Юзбашев М. М.** Общая теория статистики. – М. : Финансы и статистика, 2004. – 656 с.

REFERENCE

1. **Afanas'ev V., Yuzbashev M. M., Gulyaeva T. N.** Ekonometrika. – M. : Finansy i statistika, 2005. – 256 p.

2. **Vasil'eva E. K., Yuzbashev M. M.** Vyborochnyy metod v sotsial'no-ekonomicheskoy statistike: uchebnoe posobie (GRIF). – M. : Infra-M, 2010. – 256 p.

3. **Eliseeva I. I., Yuzbashev M. M.** Obshchaya teoriya statistiki. – M. : Finansy i statistika, 2004. – 656 p.

Малова Наталья Николаевна, *канд. экон. наук, доцент*
кафедры «Вычислительная техника и прикладная математика»
Тел. 8-909-967-53-75
E-mail: malova@bk.ru
141406, Московская область, г. Химки, ул. Совхозная д. 9, кв. 80