УДК 551.509.314: 551.557.21."321/322" (574)

ПРИМЕНЕНИЕ МЕТОДА СКЛАДНОГО НОЖА И БУТСТРЕП – ПРОЦЕДУРЫ В СТАТИСТИЧЕСКИХ МОДЕЛЯХ ПРОГНОЗА

Канд. физ.- мат. наук Е.В. Боголюбова

Статья знакомит с новыми подходами к обработке массивов статистических данных, а также с использованием метода складного ножа и бутстреп-процедуры в регрессионных моделях прогноза погоды на примере прогноза месячных сумм осадков на территории Казахстана. Эта процедура позволяет уменьшить смещение и повысить устойчивость статистических моделей.

Статистические модели прогнозов, применяемые в метеорологии, чаще всего бывают двух типов: модели, основанные на классификации, и модели, основанные на регрессии. Очень часто в эти модели вводится предварительно «сжатая» информация. «Сжимают» её обычно с помощью метода главных компонент, разложения по полиномам Чебышева либо с помощью других методов аналитического представления метеорологических полей.

Многие ученые наиболее результативными и перспективными считали методы экстраполяции периодических составляющих и методы линейной и нелинейной регрессии. Однако модели с использованием только периодических составляющих оказались неустойчивыми. Среди моделей, наиболее часто применяемых в метеорологических исследованиях, выделяются линейные регрессионные модели и их частные случаи – множественная линейная регрессия, регрессия с использованием главных компонент устойчивая регрессия («складной нож»), часто с применением пошаговой процедуры.

Классическая теория наименьших квадратов (использование принципа Лежандра) накладывает на имеющиеся данные ряд ограничений. Обычные предположения относительно ошибок состоят в признании ошибок независимыми, имеющими нулевые средние, постоянную дисперсию и подчиняющимися нормальному распределению. Все это предполагает нормальность предиктанта и несмещённость оценок регрессии. То есть в естественных случаях (например, с осадками), когда нормальность предиктанта не имеет места, мы имеем дело с нарушениями исходных предпосылок. А нарушение исходных предпосылок в методе наименьших

квадратов может привести к неустойчивости результатов и смещению регрессионных оценок.

Смещение возникает в основном по двум причинам:

- из-за «неадекватности» модели; в частности, когда не все предикторы учтены;
- из-за несимметричности ошибок в модели.

Смещение из-за неполноты модели может быть уменьшено введением дополнительных предикторов, но здесь возникают определенные трудности с их выбором и увеличением дисперсии предсказания с ростом числа предикторов.

Что касается второй причины, то, по мнению Тьюки [2], смещение может быть уменьшено применением метода «складного ножа» (jackknife). Иначе этот метод иногда называют методом расщепления выборки. Тьюки показал, что для оценки Y_n параметра θ, смещение которой допускает асимптотическое разложение $M(Y_n-\theta)=\frac{a_1}{n}+\frac{a_2}{n^2}+o(n^{-3})$ при $n\to\infty$, оценка, построенная по псевдозначениям «складного ножа» $Y^*=\Sigma\frac{1}{n}Y_n^*$ имеет меньшее смещение $M(Y_n-\theta)=-\frac{a}{n^2}+o(n^{-3})$ при $n\to\infty$, где Y_n – оценка параметра θ, a_1, a_2 – коэффициенты, выражение $o(n^{-3})$ означает величину порядка $\frac{1}{n_3}$.

Таким образом, в случае множественной регрессии с фиксированным значением переменных главный член смещения из-за несимметричности ошибок становится порядка $o(\frac{p^{\frac{3}{2}}}{n^2})$, вместо $o(\frac{p^{\frac{3}{2}}}{n})$, что может привести к большей устойчивости статистической модели прогноза.

Метод «складного ножа» Кенуя-Тьюки представляет собой привлекательный непараметрический прием для проверки нулевых гипотез о том, что распределение некоторой статистики симметрично относительно заданной точки, и для оценивания смещения и дисперсии изучаемой статистики. Обобщением метода «складного ножа» является бутстреп-процедура. Бутстреп отличается от многих общеизвестных методов тем, что он предполагает

многократную обработку различных частей одних и тех же данных и сопоставление полученных результатов [3]. Этот подход в свою очередь возник под влиянием идей Р. Фишера, пропагандирующего концепцию максимального правдоподобия. Реализации этого подхода мешают три обстоятельства [3]. Это возможное смещение на конечных, часто не слишком большого объема выборках, потребность в знании вида закона распределения случайных величин и, в некоторых случаях, вычислительные трудности.

Смещение может возникать по разным причинам. Источником его может быть ошибка измерения, возникающая из-за различий в измерительных приборах или навыков измерителя. Для исправления этого используют разные приемы градуирования и вычисления поправочных коэффициентов. Есть другой источник – «неадекватность модели», то есть смещение, обусловленное моделью, формулами, по которым производится расчет и по которым вычисляются статистические характеристики.

Если, например, мы считаем, что имеет место нормальное распределение, то нас не должно удивлять возникновение неточностей из-за смещения, поскольку распределение может быть не нормальным. Первая возможность борьбы с таким смещением – это знать фактическую модель. Это прием параметрической статистики, методы которой используются чаще всего. Вторая возможность состоит в применении методов непараметрической статистики и выбора модели, для которой результаты слабо зависят от действительной ситуации. Это методы робастного оценивания и рандомизированные процедуры.

Развитие вычислительной техники позволило осуществить многие вычислительные процедуры, позволяющие значительно снизить выборочное смещение. Это и была реализация разработанного в 1949 г. М. Кенуем и позднее усовершенствованного Дж. Тьюки метода «складного ножа». Бутстреп-метод («bootstrap») был предложен как обобщение метода «складного ножа». Дело в том, что формирование подвыборок в методе складного ножа, означает выбор без возвращения из имеющейся совокупности, и Б. Эфрон предложил пользоваться выбором с возвращением [3]. Тогда формально сохраняются все степени свободы на каждом этапе обработки данных.

Бутстреп-процедура не требует информации о виде закона распределения исследуемой случайной величины и именно в этом смысле может быть рассмотрена как непараметрическая.

По мнению Ю. Адлера и Ю.А. Кошевника [3] бутстреп, метод складного ножа и процедуры перепроверки тесно связаны, и не стоит их

строго различать. Например, в работах Б. Эфрона [3] они рассматриваются вместе. По мнению тех же авторов, эти методы появились для того «чтобы бороться со смещением, обусловленным выборкой. Затем выяснилось, что бутстреп стоит использовать для оценки выборочной дисперсии. Ну а от дисперсии рукой подать до доверительных границ и проверки гипотез. Таким образом, это универсальный метод».

Складной нож, бутстреп можно использовать при решении любых статистических задач: для проверки гипотез о законах распределения, в регрессионных задачах, в дисперсионном анализе, при многомерных задачах классификации. Метод вполне возможно применять и при решении сложной задачи предсказания. На основе обучающею множества можно построить эффективное правило предсказания с помощью регрессионных и других статистических схем прогноза, которые в сочетании с концепциями MOS и PP могут значительно увеличить качество прогноза локальных характеристик погоды. Б. Эфрон предлагает использовать бутстреп-процедуры при оценивании доли ошибок, например, в дискриминантном анализе [3].

В настоящей работе этот метод был использован для решения задачи предсказания с помощью линейной множественной регрессии месячных сумм осадков по сельскохозяйственным районам Казахстана. Окончательный прогноз давался по среднему «складного ножа». Изменчивость за счет изменения подвыборок оценивалась дисперсией «складного ножа», которая представляет собой оценку разброса прогноза относительно математического ожидания самого прогноза.

Так как бутстреп-процедура не требует информации о виде закона распределения, а месячные суммы осадков, являющиеся предиктантом, чаще всего распределены не нормально, использование подобной методики является правомерным. Идея метода состоит в оценивании статистики Y (месячной суммы осадков) по подвыборкам, образованным из исходной выборки. В оценку входит «межвыборочная изменчивость», влияние которой «усредняется» по подвыборкам, что приводит, по мнению Тьюки, к большей устойчивости предиктанта-прогнозируемой величины.

Опишем метод подробнее. Оценивается статистика Y (предиктант, представляющий собой месячную сумму осадков, осредненную по территории определенного района) по исходной выборке из n элементов. Пусть $Y_{общ.}$ соответствует оценке для сложной статистики, полученной по всей исходной выборке из n наблюдений. Y^j – аналогичный результат,

полученный после отбрасывание j -ого наблюдения по (n–1)-ому наблюдению. Введем псевдозначения следующим образом:

$$Y_j^* = n \cdot Y_{общ.} - (n-1) \cdot Y^j . \quad (1)$$

Теперь эти псевдозначения играют ту же роль, что и исходные данные для получения оценки Y .

Среднее складного ножа Y^* и оценка дисперсии S^{2*} даются выражениями:

$$Y^* = \frac{1}{n}(Y_1^* + Y_2^* + Y_3^* + ... + Y_n^*), \quad (2)$$

$$S^2 = \frac{\sum Y_j^2 - \frac{1}{n}\sum (Y_j^*)^2}{n-1}, \quad (3)$$

$$S^{2*} = \frac{S^2}{n}. \quad (4)$$

Как было отмечено выше, для оценки Y_n параметра θ, смещение которой допускает асимптотическое разложение $M(Y_n - \theta) = \frac{a_1}{n} + \frac{a_2}{n^2} + o(n^{-3})$ при $n \to \infty$, оценка, построенная по псевдозначениям, имеет меньшее смещение [2].

Пусть Y_n – оценка параметра θ, вычисленная по n наблюдением Y_1, Y_2, Y_3 ... Y_n, а $Y_{n-1\,j}$ – оценка той же формы, что и Y_n, но вычисленная по (n-1) наблюдениям, т.е. после отбрасывание Y_j. Обозначим за $\overline{Y}_{n-1}$ среднее по всем $Y_{n-1\,j}$, где j = 1, 2, 3, … n, предположим, что для всех достаточно больших n, в частности для $m = n-1$ и $m = n$.

$$M(Y_n) = \theta + \frac{a_1(\theta)}{m} + \frac{a_2(\theta)}{m^2} + o\left(\frac{1}{m^3}\right).$$

Тогда $M(Y_{n-1}) = \theta + \frac{a_1(\theta)}{n-1} + \frac{a_2(\theta)}{(n-1)^2} + o\left(\frac{1}{n^3}\right),$

$$M(Y_n) = \theta + \frac{a_1(\theta)}{n} + \frac{a_2(\theta)}{n^2} + o\left(\frac{1}{n^3}\right).$$

Из двух последних равенств вытекает, что можно найти такую линейную комбинацию статистик $\overline{Y}_{n-1}$, Y_n которая будет иметь смещение порядка $\frac{1}{n^2}$ т.е. можно исключить член порядка $o\left(\frac{1}{n}\right)$. Действительно

$$Y_n^* = n \cdot Y_n - (n-1)\overline{Y}_{n-1} = n \cdot Y_n - \frac{n-1}{n}\sum Y_{n-1j} \qquad (5)$$

имеет математическое ожидание $M(Y_n^*) = \theta + o(\frac{1}{n^2}) + o\frac{1}{n^3}$.

Поэтому в случае множественной регрессии с фиксированным значением числа переменных *p* главный член смещения из-за несимметричности ошибок становится $o\left(\frac{p^{\frac{3}{2}}}{n^2}\right)$ вместо $o\left(\frac{p^{\frac{3}{2}}}{n}\right)$. Дж. Тьюки [2] указывает на большую устойчивость оценок, рассчитанных с помощью псевдозначений, по сравнению с оценками, полученными по исходной выборке. Выборочная изменчивость полученной статистики оценивается величиной S^{2*}, которая, как показал Тьюки, служит приближением дисперсии оценки статистики Y. Таким образом, применяя складной нож, бутстреп – процедуру мы имеем следующие преимущества:

1. меньшее смещение по сравнению с исходной оценкой;
2. большую устойчивость, что позволяет использовать анализ остатков для корреляции предсказания;
3. возможность оценивать дисперсию предсказания в условиях, далеких от идеальных.

В настоящей работе складной нож применялся к прогностической модели с помощью уравнения линейной множественной регрессии, где отбор предикторов проводился на основе дальних корреляционных связей осадков с характеристиками полей геопотенциала поверхностей H_{500}, H_{700} и H_{1000}^{500} с помощью пошаговой процедуры.

Прогноз строился по следующей схеме. Для каждой из 12 областей, по которым дается прогноз, по исходной выборке длиною *n* = 30 оценивалось уравнение регрессии

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_k X_k,$$

рассчитывались коэффициенты и строился прогноз по исходному набору предикторов, т.е. $Y_{общ} = a_0 + a_1 X_1 + \cdots + a_k X_k$.

Для построения *j*–го псевдозначения из исходной выборки выбрасывалось *j*–ое наблюдение и по оставшейся выборке из $n - 1$ элементов

$$x^{(1)}_1, x^{(1)}_2, \ldots, x^{(1)}_r ; y^{(1)}$$
$$x^{(2)}_1, x^{(2)}_2 \ldots, x^{(2)}_k ; y^{(2)}$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$x^{(j-1)}_1, x^{(j-1)}_2, \ldots, x^{(j-1)}_r ; y^{(j-1)}$$
$$x^{(j+1)}_1, x^{(j+1)}_2, \ldots, x^{(j+1)}_r ; y^{(j+1)}$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$x^{(n)}_1, x^{(n)}_2, \ldots, x^{(n)}_k ; y^{(n)}$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

строилось уравнение

$$Y^j = a^j_o + a^j_1 X_1 + \ldots + a^j_k X_k ,$$

вычислялись коэффициенты и рассчитывалась оценка прогноза

$$Y^j = a^j_o + a^j_1 x_1 + \ldots + a^j_k x . \qquad (7)$$

Затем определялись псевдозначения по формуле

$$Y^*_j = nY_{общ} - (n-1)Y^j . \qquad (8)$$

Далее рассчитывались средняя и дисперсия складного ножа по формулам (2), (3), (4). Окончательный прогноз есть среднее складного ножа

$$\overline{Y}^* = \frac{\Sigma Y^*_j}{n} . \qquad (9)$$

Для оценки качества прогноза был реализован метод скользящего контроля, который представляет собой универсальный метод оценивания статистических моделей в тех случаях, когда выборка недостаточно велика и позволяет получать оценки на числе случаев, равном объему исходной выборки. Скользящий контроль состоит в поочередном исключении ситуации из выборки и проверке на этой ситуации качества работы модели, построенной по оставшимся ситуациями. Суммирование ошибок, полученных с помощью скользящего контроля по всем ситуациям, дает возможность оценить среднюю ошибку на независимом материале.

Метод был реализован следующим образом. Из исходной выборки исключалось первое наблюдение, которое считалось независимым для прогноза. Оценки с помощью складного ножа получались на оставшейся выборке. Затем первое наблюдение возвращалось обратно, отбиралось

следующее наблюдение, оценки получались на оставшейся выборке, отличной от первой, и процесс продолжался далее. Модель каждый раз менялась и оценки получались каждый раз на новой выборке.

Если оценки устойчивы, то прогноз является достаточно удовлетворительным. Если они неустойчивы, то и оценки прогноза вряд ли можно считать хорошими. Изменчивость за счет изменения подвыборок оценивается дисперсией складного ножа. Если дисперсия псевдозначений велика, то метод скорее всего неустойчив.

Прогноз давался в виде числа, затем оценивался по принадлежности к трем равновероятным градациями («ниже нормы», «норма», «выше нормы»), полученным с помощью эмпирических функций распределения.

В процессе использования скользящего контроля были получены оценки качества прогноза по критерию p (критерию принадлежности к определенной градации), критерию ρ (критерию оценки знака аномалии) и по коэффициентам корреляции r между прогностическим и фактическим значениями месячных сумм осадков. Это обычные формулы, используемые при оценке месячных прогнозов погоды, известные всем специалистам в этой области. Оценка проводилась по каждому из 12 выбранных административных районов для каждого месяца теплого времени года (апрель...октябрь) и в целом по территории.

Приведем средние оценки по этим критериям за весь период. В табл. первыми указаны бывшие названия областей.

Таблица

Оправдываемость прогноза месячных сумм осадков с помощью скользящего контроля по критериям p, ρ, r и в среднем за весь теплый период

Область	Оправдываемость прогноза по		
	p %	ρ	r
Гурьевская (Атырауская)	76	0,35	0,45
Уральская (север)	79	0,60	0,60
Уральская (юг)	73	0,49	0,51
Актюбинская	75	0,47	0,54
Семипалатинская (север)	73	0,39	0,40
Семипалатинская (центр)	77	0,42	0,62
Северо – Казахстанская	77	0,39	0,50
Кокчетавская	78	0,38	0,51
Целиноградская (Акмолинская)	68	0,49	0,39
Павлодарская	76	0,48	0,57
Карагандинская	77	0,48	0,65
Кустанайская	73	0,41	0,48
Среднее по территории	75	0,45	0,52

На независимом материале для 60 случаев оценка по p составила 78 %, а по $\rho = 0{,}53$.

Все это позволяет придти к следующему мнению. При физически обоснованном выборе предикторов на основании статистически значимых корреляционных связей «jackknife» и «bootstrap» – процедуры можно эффективно использовать в регрессионных и других статистических схемах для предсказания отдельных характеристик погоды.

СПИСОК ЛИТЕРАТУРЫ

1. Боголюбова Е.В. Прогноз месячной суммы осадков в весенне-летний, период по сельскохозяйственным районам Казахстана: Автореф. дис. ... канд. физ.-мат. наук. – М., 1985, – 25 с.

2. Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия. Вып. 1, 2 - М.: Финансы и статистика, 1982. – 319 с., 239 с.

3. Эфрон Б. Нетрадиционные методы многомерного статистического анализа. – М.: Финансы и статистика, 1988. – 263с.

Казахский национальный университет им. аль-Фараби

БОЛЖАУДЫҢ СТАТИСТИКАЛЫҚ ҮЛГІСІНДЕ “СКЛАДНОЙ НОЖ ЖӘНЕ БУТСТРЕП-ПРОЦЕДУРА” ӘДІСІН ПАЙДАЛАНУ

Физ.-мат. ғылымд. канд. Е.В. Боголюбова

Мақалада көлемді статистикалық деректерді өңдеудің жаңа жолы Қазақстан аумағындағы жауын-шашынның айлық жиынтығын болжау мысалында ауа райын болжаудың регрессивті үлгісінде аталған әдісті пайдалану барысында көрсетілген. Мұндай процедура жылжуды бәсеңдетіп, статистикалық үлгінің тұрақтылығының артуына мүмкіндік береді.