Материалы региональной конференции (с участием иностранных ученых) "Вероятностные идеи в науке и философии". – Новосибирск: Ин-т философии и права СО РАН / НГУ. 2003. – С. 106-109.

Проблемы применения классического аппарата дисперсионного анализа в приложениях технического, экономического и естественно-научного характера[1]

Лемешко Б.Ю. Пономаренко В.М.
НГТУ, Новосибирск. E-mail: ponomarenkov@mail.ru

Возникновение дисперсионного анализа связано с решением в 20–30-х годах прошлого столетия Р. Фишером и его учениками задач анализа сельскохозяйственных данных. С тех пор аппарат дисперсионного анализа был существенно развит и широко применяется в различных областях науки, техники и сельского хозяйства. Проверка гипотез в классическом дисперсионном анализе базируется на ряде предположений, в том числе, на предположении о нормальном распределении ошибок наблюдений. В рамках этих допущений оказалось возможным аналитически вывести предельные распределения статистик. И, хотя в реальных задачах далеко не всегда выполняются постулируемые условия, вследствие отсутствия альтернативы, применяют классический аппарат дисперсионного анализа. Такая практика может приводить к некорректным статистическим выводам.

В данной работе методами статистического моделирования исследуется, что происходит с предельным распределением статистики отношения правдоподобия при нарушении предположения о нормальности ошибок наблюдений при использовании некоторых моделей со сбалансированными и несбалансированными планами.

Рассматриваемая модель отклика имеет вид:

$$Y = X\theta + e, \tag{1}$$

где Y – вектор наблюдений размерности $(n\times 1)$, X – матрица планирования размерности $(n\times m)$, $r = rg(X)$ ранг X, θ - вектор оцениваемых параметров модели размерности $(m\times 1)$, e – случайный вектор ошибок наблюдения размерности $(n\times 1)$. Компоненты $(e_1,\ldots,e_n)$ вектора e предполагаются независимыми случайными величинами, одинаково распределенными с общей функцией распределения, имеющей нулевое математическое ожидание и некоторую истинную дисперсию σ^2.

В самом общем виде линейную гипотезу относительно параметров модели можно представить следующим образом:

$$H: K^T\hat{\theta} = b, \tag{2}$$

где K^T - известная матрица размерности $(k\times m)$, $rg(K) = k \le m$; b - заданный вектор размерности $(k\times 1)$, $\hat{\theta}$ - оценка вектора параметров модели θ. Для проверки гипотез вида (2) в дисперсионном анализе используется статистика вида:

$$Q = \frac{\left(K^T\hat{\theta} - b\right)^T\left(K^T G K\right)^{-1}\left(K^T\hat{\theta} - b\right)}{\left(Y - X\hat{\theta}\right)^T\left(Y - X\hat{\theta}\right)}\frac{n-r}{k}, \tag{3}$$

называемая статистикой отношения правдоподобия, где G - матрица, обобщенно-обратная к матрице $X^T X$.

Оценивание параметров модели θ осуществляется методом наименьших квадратов (МНК). Дисперсия ошибок наблюдения σ^2, хотя и может рассматриваться как один из параметров модели, при проводимых исследованиях не оценивается наравне с параметрами модели θ, и относительно дисперсии не проверяется никаких гипотез. Шеффе в [1] называет такую ситуацию проверкой «гипотез о среднем». Однако оценка дисперсии ошибок наблюдений

$$s^2 = \frac{\left(Y - X\hat{\theta}\right)^T\left(Y - X\hat{\theta}\right)}{n-r}$$

включена непосредственно в формулу (3) статистики отношения правдоподобия.

Моделирование данных, использовавшихся при проведении исследований, осуществлялось следующим образом. Создавался набор из десяти незашумленных моделей. Все незашумленные модели в рамках одного набора имели одинаковые размерность, матрицу планирования, вид проверяемой

[1] Работа выполнена при финансовой поддержке Минобразования России (проект № Т02-3.3-3356)

гипотезы, и семейство распределений, которому подчиняются ошибки наблюдений. Незашумленные модели в рамках одного набора различаются задаваемыми значениями параметров модели θ. Это позволяет говорить о том, что исследования в рамках одного набора моделей проводились в идентичных условиях. Матрица планирования X соответствует модели для полного факторного эксперимента с повторными наблюдениями. Случай одинакового числа повторов во всех точках плана соответствует сбалансированному плану, разного – несбалансированному плану. Далее, случайным образом для каждой незашумленной модели одного набора генерировался вектор «истинных» параметров модели $\theta_{ист}$. На основании этого вектора и матрицы планирования X формировался «истинный» (т.е. незашумленный) вектор откликов $Y_{ист}$:

$$Y_{ист} = X\theta_{ист}.$$

В наших исследованиях использовалось понятие уровня шума ρ, в соответствии с которым дисперсия ошибки σ^2 вычислялась по формуле:

$$\sigma^2 = \frac{\rho}{100}c^2, \quad \text{где } c^2 = \frac{1}{n}\sum_{i=1}^{n}((Y_{ист})_i - \overline{Y}_{ист})^2, \quad \overline{Y}_{ист} = \frac{1}{n}\sum_{i=1}^{n}\left|(Y_{ист})_i\right|.$$

Для всех моделей $\rho = 10\%$.

Для каждой незашумленной модели генерировалась серия из 2000 экспериментов, отличающихся друг от друга реализациями случайных величин, являющихся компонентами вектора ошибок наблюдений e. В результате каждого эксперимента вычислялось одно значение статистики (3). Таким образом, для каждой незашумленной модели формировалась выборка из 2000 значений статистик. По каждой выборке статистик проводилась проверка согласия полученного эмпирического распределения с распределением, которому подчиняется статистика (3) в случае нормального распределения ошибок наблюдений. В [1] показано, что это F-распределение Фишера со степенями свободы *k* и *n-r*. Распределение Фишера $F(k, n-r)$ является частным случаем Бета-распределения II рода [2]:

$$F(k, n-g) = Be_{II}(0,\ (n-g)/k;\ k/2,\ (n-g)/2). \qquad (4)$$

Применительно к показателям используемой нами модели (1), статистика (3) в нормальном случае будет иметь распределение со степенями свободы $rg(K)$ и $n - rg(X)$.

Если не наблюдалось согласия эмпирического распределения с соответствующим распределением Фишера, то выборки анализировались на предмет принадлежности к некоторому классу теоретических распределений и оценивались неизвестные параметры предполагаемого распределения. Во всех случаях оценивание параметров распределения осуществлялось методом максимального правдоподобия. В первом случае проверялась простая гипотеза о согласии, во втором случае – сложная гипотеза. В обоих случаях гипотеза о согласии отвергалась, если среднее арифметическое достигнутых уровней значимости при проверке гипотезы с помощью критериев отношения правдоподобия, χ^2 Пирсона, Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса было меньше задаваемого уровня ошибки 1-го рода $\alpha = 0.05$. Когда усредненный по критериям достигаемый уровень значимости значительно больше 0.05, речь идет о хорошем согласии.

Как при проверке согласия с известным распределением, так и при построении моделей распределений использовался программная система ISW (Интервальная статистика), развиваемая на кафедре прикладной математики НГТУ [3].

Приближение предельного закона распределения статистики формировалась как усредненное по параметрам распределение, полученное на основании десяти найденных распределений (по десяти выборкам, соответствующих десяти моделям). Согласие полученного приближения проверялось с эмпирическими распределениями того набора моделей, который участвовал в формировании приближения, а также дополнительного набора. Во всех рассмотренных случаях наблюдалось хорошее согласие.

Исследования предельных распределений статистики (3) проводились для следующих распределений ошибок наблюдений: нормального, Коши, экспоненциального семейства распределений с параметром формы λ, принимающим значения 0.3, 0.5, 1, 3, 5, 10, обозначаемого далее как De(λ), распределений минимального и максимального значения.

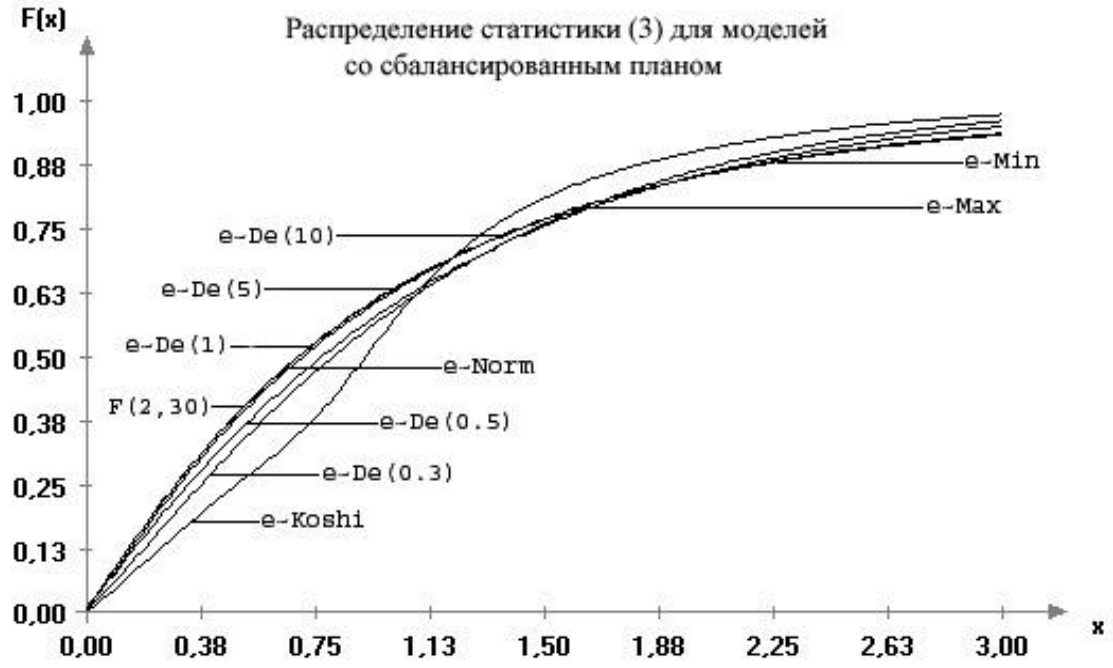
Исследования проводились на нескольких наборах сбалансированных моделей и на одном наборе несбалансированных моделей. Размерности моделей были одинаковы для всех наборов: *n*=36, *m*=8, *r*=6, *k*=2. Наборы сбалансированных моделей различались только множеством значений параметров модели

θ. Набор несбалансированных моделей отличался от одного из наборов сбалансированных моделей только матрицей планирования X: в точках плана, соответствующих нахождению первого фактора на первом уровне повторов не было (т.е. имелось только по одному наблюдению); в точках плана, соответствующих нахождению первого фактора на последнем, третьем уровне число повторов увеличилось с 2-х до четырех; в точках плана, соответствующих нахождению первого фактора на втором уровне число повторов сохранилось равным двум.

Распределение статистики (3) (в случае нормально распределенных ошибок наблюдения) для всех выше описанных наборов моделей – $F(2,30)$.

При проведении исследований изучались следующие вопросы. Во-первых, насколько сильно и каким образом получаемые для разных распределений ошибок приближения предельного распределения статистики отклоняются от предельного распределения статистики в нормальном случае (от F-распределения Фишера). Во-вторых, насколько сильным оказывается разброс получаемых эмпирических распределений вокруг вычисляемого на их основе приближения предельного закона. И, в-третьих, каково влияние несбалансированности плана.

В сбалансированном случае были получены следующие результаты. Во-первых, значимое отклонение получаемого приближения от F-распределения Фишера наблюдается только в случае распределения ошибок по Коши, De(0.3) и De(0.5). Для этих распределений ошибок были получены следующие приближения предельных распределений статистики: в случае распределения ошибок наблюдений по закону Коши – смесь распределений максимального значения (с параметром сдвига 0.949, масштаба 0.209) и Вейбулла (с параметром формы 1.157, сдвига 0.0001, масштаба 1.085) с параметром смеси, равным 0.253; в случае распределения ошибок наблюдений по закону De(0.3) – распределение Вейбулла (с параметром формы 1.176, сдвига -0.007, масштаба 1.123); в случае распределения ошибок наблюдений по закону De(0.5) – распределение Вейбулла (с параметром формы 1.0798, сдвига -0.0083, масштаба 1.1039).

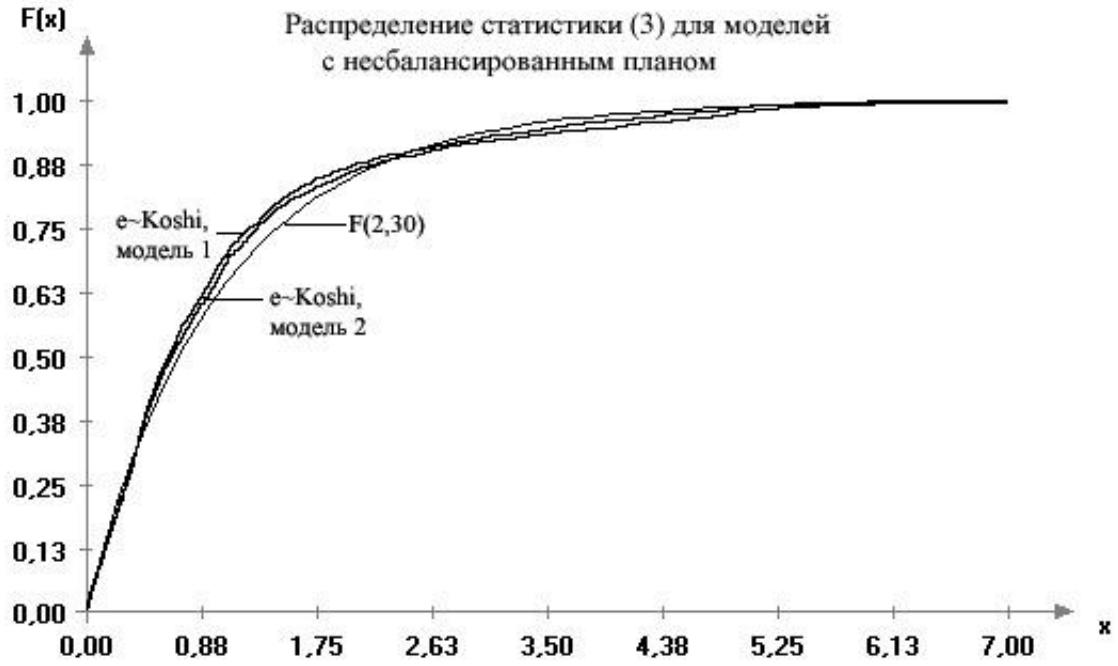


В остальных случаях, в том числе и в случае несимметричных законов распределения ошибок: максимального и минимального значений, наблюдалось согласие эмпирического закона распределения статистики с F-распределением Фишера, соответствующим предельному распределению статистики в нормальном случае. Полученные результаты иллюстрирует первый рисунок (надпись $e \sim \ldots$ означает, что указанное приближение предельного распределения статистики (3) было получено в случае распределения ошибок наблюдения по обозначенному закону распределения).

Во-вторых, во всех рассмотренных случаях разброс эмпирических распределений статистики вокруг полученного на их основе приближения весьма незначителен.

В несбалансированном случае отклонения распределений статистики от распределения Фишера наблюдается для тех же распределений ошибок, но характер отклонения меняется, и степень разброса

эмпирических распределений немного увеличивается. Модели распределения статистики построить пока не удалось, а характер отклонения на примере эмпирических распределений статистик двух первых моделей из набора, обозначенных как «модель 1» и «модель 2», иллюстрирует второй рисунок.



Заключение. В целом можно сделать следующий вывод: предельное распределение статистики (3) довольно устойчиво к изменению закона распределения ошибок наблюдений в случае оценивания параметров модели по МНК. Значимые отклонения эмпирических распределений от F -распределения Фишера как для сбалансированных, так и для не сбалансированных планов наблюдались только при распределении ошибки наблюдений по Коши, De(0.3) и De(0.5). При указанных распределениях ошибок остается проблема построения приближений предельных распределений статистик, удобных для практического использования. Необходимо также провести более углубленный анализ влияния степени и различных типов несбалансированности плана на получаемые результаты.

Таким образом, при использовании процедур классического дисперсионного анализа полностью игнорировать проверку предположения о нормальности все же нельзя.

1. Шеффе Г. Дисперсионный анализ. М.: Физматгиз, 1980.
2. Губарев В.В. Вероятностные модели: Справочник в 2-х ч. // Новосиб. электротехн. ин-т. - Новосибирск, 1992. - 422 с.
3. Лемешко Б.Ю., Постовалов С.Н. Система статистического анализа наблюдений и исследования статистических закономерностей // Сб. "Моделирование, автоматизация и оптимизация наукоемких технологий". - Новосибирск: изд-во НГТУ, 2000. - С. 44-46.