УДК 519.281: 528.11: 31(075.8)

Джунь Йосип Володимирович, д.ф.-м.н., професор, завідувач кафедри математичного моделювання (Міжнародний економіко-гуманітарний університет імені академіка Степана Дем'янчука, м. Рівне)

СУЧАСНІ ТЕНДЕНЦІЇ В МАТЕМАТИЧНОМУ МОДЕЛЮВАННІ І ІНТЕЛЕКТУАЛЬНОМУ АНАЛІЗІ ДАНИХ

Знаменитий англійський математик і статистик Карл Пірсон писав, що будь-яке наукове досягнення не є остаточним, а являє собою лише найімовірніший висновок, який отриманий на основі даних, які є у автора. Більш об'ємні вибірки або більш досконалі: аналіз, експеримент чи спостереження, приводять до нових формул і теорій [1]. Це твердження Пірсона і нині є актуальним по відношенню до всіх наук і особливо для таких важливих розділів математичної статистики як моделювання і інтелектуальний аналіз даних. Що ж саме, які процеси спричинили еволюцію підходів в зазначених вище розділах математичної статистики? Постараємося коротко, але ґрунтовно відповісти на це важливе питання.

По – перше, і це є дуже важливий фактор,– істотно змінились уявлення про ті фундаментальні аксіоми, які є наріжними в моделюванні. Як відомо, головним інструментом математичного моделювання є нині метод найменших квадратів (МНК), який запропонував і найбільш повно виклав геніальний німецький математик К. Ф, Гаус ще в 1809 році у своїх знаменитих мемуарах [2]. Перше наукове застосування МНК пов'язано з пошуком малої планети Церери, яка була відкрита Д. Піаці в 1801 році, але потім загублена. Гаус на основі МНК точно вказав координати небесної сфери де ця планета мала появитися. Там її і знайшли, так би мовити «на кінчику пера». Це був блискучий успіх Гауса, який приніс йому світову славу і підтвердив надійність його методу моделювання.

МНК ґрунтується на двох фундаментальних аксіомах: 1) випадкові похибки спостережень ϑ підкоряються закону $e^{-n\cdot\vartheta^2}$, тобто, нормальному розподілу; 2) в результатах спостережень відсутні систематичні похибки. Це і є основоположні принципи класичного моделювання, яке більше ніж 130 років успішно застосовувалось в найрізноманітніших галузях науки і техніки, аж до того часу поки не запрацювала в Грінвічській обсерваторії (Англія) перша автоматизована система спостережень. Це була знаменита фотографічна зенітна труба Куксона, яка плавала на ртуті і була встановлена для автоматизованого слідкування за зміною широти. За період 1932–1936 рр. на цій трубі було зроблено 4982 спостережень, які у королівських астрономів Х. Р. Хюльме і Л. С. Т. Сімса викликали вкрай велике здивування.

Виявилося, що похибки цих спостережень аж ніяк не підкорялись закону $e^{-n\cdot\vartheta^2}$: вони мали фантастично великий ексцес ε = + 6.00 ± 0.06, в той час як у нормального закону ε=0! Число похибок $|\vartheta|$, більших ніж 3σ склало 453 спостереження, тобто 9,1 % замість необхідних 0,27 % по Гаусу. Тоді ж було помічено, що похибки невеликих за обсягом спостережень n < 400– 500, які, як правило, проводяться вручну, задовільно описуються законом $e^{-n\cdot\vartheta^2}$. При автоматизованих спостереженнях, які мають величезні обсяги, ми ніколи не спостерігаємо закону Гауса. Практично це означає, що МНК при обробці спостережень значного обсягу не може забезпечити ефективність оцінювання, «Вирішальним питанням в комбінації спостережень – як зазначив в [3] знаменитий вчений Кембриджського університету Г. Джеффріс – є знання того, чи дійсно розподіл похибок слідує нормальному закону, якщо це не так, то потрібно придумати інші методи, властиві даному закону» [3]. А це значить, що і програмне забезпечення в цьому разі не повинно ґрунтуватись виключно на МНК Гауса, який лише і вивчають у вузах. Створення автоматизованих систем є найбільш істотним і важливим викликом сучасної епохи. Він означає невпинне зростання обсягів вимірювань внаслідок їх автоматизації і комп'ютеризації. Розглядаючи розподіли спостережень великих обсягів Джеффріс зробив впевнений висновок, що при числі спостережень n > 500 гіпотеза нормальності і практично і теоретично є неспроможною, оскільки в даному випадку похибки підкоряються зовсім іншому закону, ніж закон Гауса, а саме – розподілу Пірсона *VII* типу з діагональною інформаційною матрицею Фішера:

$$f(x)=\left[\sigma\sqrt{2m-1}B\left(m+\frac{1}{2},\frac{1}{2}\right)\right]^{-1}\left[1+\frac{0,5}{M}\left(\frac{x-a}{\sigma}\right)^2\right]^{-m}, \qquad (1)$$

де a, σ – математичне сподівання і параметр розсіювання, а m – ключовий параметр закону (1), що є мірою його відхилення від нормального закону; $B(z, w)$ – бета – функція; $M = (m - 0,5)^3 \cdot m^{-2}$.

Тобто, нині необхідно освоювати обробку спостережень при негаусових розподілах їх похибок – що і є новою і найважливішою тенденцією в сучасному моделюванні.

По – друге: використання форми (1), дозволяє вирішити проблему контролю відсутності невипадкових, тобто корельованих похибок в результатах спостережень. Джеффріс показав в [4], що для повністю незалежних випадкових похибок параметр m має бути в межах:

$$3 \le m \le 5 \qquad (2)$$

Це означає, що за допомогою відношення (2) дослідник може контролювати спроможність моделі, тобто, її якість за допомогою параметра *m*, отриманого для її залишкових похибок. Це є другою, важливою тенденцією в сучасному моделюванні, яка вперше створює прецедент його діагностики. До цього часу, в жодному із програмних продуктів відсутні процедури діагностики моделювання, хоч це вкрай важливо.

І, на кінець, третя особливість сучасних методів моделювання в тому, що форма (1) дозволяє дуже просто отримати вагову функцію похибок ϑ:

$$P(\vartheta) = \left[\left(\frac{m-0,5}{m} \right)^3 \sigma^2 + \frac{\vartheta^2}{2m} \right]^{-1}, \qquad (3)$$

де $\vartheta = x - a$.

Формула (3) забезпечує процедури некласичного МНК, які дозволяють отримати ефективні оцінки регресорів і їх стандарти. Ці процедури ретельно описані в монографії [4], розробленій на кафедрі математичного моделювання факультету кібернетики Міжнародного економіко-гуманітарного університету імені академіка Степана Дем'янчука і виданій у святковому варіанті в 2015 році. Ця монографія отримала біля десяти схвальних відгуків фахівців в галузі аналізу даних як зарубіжних, так і вітчизняних. Наприклад, з Одеської академії будівництва і архітектури монографія [4] віднесена до однієї з найвизначніших в XXI столітті.

Висновок: для ознайомлення з сучасними тенденціями в математичному моделюванні, його ідеями і особливостями, потрібно звернутись до фундаментальної праці Г. Джеффріса [3] і монографії [4]. Ці ідеї, їх алгоритм необхідно використовувати в учбовому процесі, знайомити з ними студентів, аспірантів, всіх, хто займається сучасними дослідженнями і аналізом даних.

Нами озвучені останні новини в галузі математичного моделювання і інтелектуального аналізу даних. Чи можна ігнорувати зроблені висновки і використовувати тривіальні процедури, розроблені більше ніж 200 років тому. Звичайно можна! Але тоді ми випадаємо за межі сучасної науки і межі високих досягнень кембриджської наукової школи аналізу даних. А це означає ризик назавжди лишитись в обіймах рутинного обскурантизму, закривати перспективи якісних змін у способах вдосконалення своєї наукової роботи, моделювання якості спостережень, алгоритмі програмування.

1. Pearson K. Grammar of Science. New York: Dover Publications, 2004–642 p.

2. Gauss C. F. Theoria motus corporum colestium in sectionibus conicis Solem ambientium. Hamburgi, 1809.

3. Jeffreys H. Theory of Probability. Clarendon Press, 1998 – 459 p.

4. Джунь И. В. Неклассическая теория погрешностей измерений. Ровно: Естеро. 2015. 168 с.