УДК 004.82

ПОДХОД К РАЗРАБОТКЕ СРЕДСТВ АНАЛИЗА ДЛЯ БИОИНФОРМАТИКИ

В. И. Хведчук, А. И. Самаха

Брестский государственный технический университет, Брест, Беларусь

As a formal model, we choose the basic elements of mathematical statistics, the prerequisites of the least squares method. The quality of the regression equation is analyzed in the following areas: checking the statistical significance of the coefficients of the regression equation, checking the overall quality of the regression equation. The Mathcad system was chosen as an instrumental system.

Введение

Отличительной чертой современного этапа развития естествознания является математизация, а использование статистических методов для проверки выдвинутых гипотез, обоснованного формирования выборок, построения математических моделей различных явлений и процессов – ее неотъемлемая часть. Практически нет такого метода статистического анализа, который не нашел бы применения в медицине [1].

1. Обзор известных решений

Известна работа [2], в которой предлагается модель для диагностики кровоизлияний, опухолей головного мозга человека, сердечных заболеваний и заболеваний щитовидной железы. Система использует нечеткую логику: фаззификатор, механизм логического вывода, базу правил и дефаззификацию для следующей модели. Модель принимает пять входных данных: белок, эритроциты, лимфоциты, нейтрофилы, эозинофилы и выдает три выходных данных: норма, кровоизлияние, опухоль головного мозга при заболевании головного мозга. В то время как при заболеваниях сердца используется только один входной сигнал: значение C.P.K.M.B, указывает на наличие заболевания сердца или нет. Аналогично, для определения заболевания щитовидной железы используются три входных сигнала: значение T-3, количество T-4, ультрачувствительный гормон (T.S.H), который выдает наличие заболевания щитовидной железы или нет. Медицинский диагноз для модели формулируются и применяются нечеткие правила с использованием моделирования в среде MATLAB. Результаты моделирования рассчитываются на основе расчетной модели. Предлагается разработать систему для повышения эффективности диагностики заболеваний, связанных с заболеваниями человека.

2. Математическая модель исследования

В качестве формальной модели выбираем базовые элементы математической статистики [3].

2.1 Предпосылки метода наименьших квадратов (МНК).

1) Математическое ожидание случайного отклонения ε_j равно 0 для всех наблюдений.

2) Постоянство дисперсии отклонений.

$D(\varepsilon_i)=D(\varepsilon_j)=a$ для любых i и j. При невыполнении этого условия имеет место гетероскедастичность.

3) Отсутствие автокорреляции для случайных отклонений ε_i и ε_j.

$$\sigma_{\varepsilon i,\varepsilon j} = \mathrm{cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} = 0, при\_ i \neq j \\ = 1, при\_ i = j \end{cases}$$

4) Случайное отклонение должно быть независимо от объясняющих переменных $o''_{ei,xi} = O$

5) Модель является линейной относительно параметров. Для множественной линейной регрессии необходимо еще 6).

6) Отсутствие мультиколлинеарности. Между объясняющими переменными отсутствует строгая линейная зависимость.

7) Ошибки Sj имеют нормальное распределение *(s* ~ N(0, a).

Это необходимо для выполнения статистических гипотез и построения интервальных оценок.

Качество уравнения регрессии анализируется по следующим направлениям.

1) Проверка статистической значимости коэффициентов уравнения регрессии. Ведется на базе статистики, имеющей распределение Стьюдента, t=bj/Sbj.

Если |t|>t(α/2,n-m-l), то bj считается статистически значимым, если bj незначим, то Xj лучше исключить из модели.

2) Проверка общего качества уравнения регрессии. Ведется на базе коэффициента детерминации

$$R^2=1-\frac{\sum e_i^2}{\sum (y_i-\bar{y})^2}$$

Могут быть неправильно специфицированные модели с высоким коэффициентом детерминации.

Проверяется также гипотеза общей значимости (о равенстве 0 всех коэффициентов регрессии):

Проверяется также равенство 0 части коэффициентов регрессии. Для модели Y=b0+blXl...bmXm рассчитывается R_1^2, для модели с k удаленными последними переменными R_2^2.

Используется статистика: $F = ((R_1^2 - R_2^2)/(1-R_1^2)) * (n-m-1)/k$

Если Fнабл>F(α ,m,n-m-l), то исключение k переменных некорректно. При этом необходимо, чтобы зависимая переменная была представлена в одной и той же форме и число наблюдений было одинаково.

Возможна проверка совпадений уравнений регрессии для отдельных групп наблюдений с использованием теста Чоу. Имеется 2 выборки объемом nl и п2, для каждой из выборок оценено уравнение регрессии $\sum e_{i.k}^2 = Sk..$

Оценивается S0 для объединенной выборки объемом nl+n2 и F=((S0-S1 - S2)/(S 1+S2))* (n 1 +n2-2m-2)/(m+1)

vl=m+l, v2=nl+n2-2m-2.

Если Fнабл<F(α, vl, v2), то уравнения регрессии для обеих выборок практически одинаковы.

2.2. Гетероскедастичность

2.2.1 Обнаружение гетероскедастичности

Тесты и критерии для обнаружения гетероскедастичности не являются однозначны-ми.

2.2.1.1 Графический анализ остатков

Строится зависимость отклонения ei (или e_i^2) от объясняющей переменной X (либо Y = Ь0 + blxl + ... + bmXm). При наличии систематических изменений отклонений от X или Y, говорят о наличии гетероскедастичности. При множественной регрессии строятся зависимости отклонений от объясняющих переменных xj или y (при линейной регрессии).

2.2.1.2 Тест ранговой корреляции Спирмена

xi и ei сортируются по возрастанию. Затем определяется коэффициент ранговой корреляции:

$r_{x.e} = 1 - 6\dfrac{\sum d_i^2}{n(n^2-1)}$, где di=r(xi)-r(ei), r(xi)-ранг xi, r(ei)-ранг ei.

Рассчитывается статистика

$$t = \frac{r_{x,e}\sqrt{n-2}}{\sqrt{1-r_{x,e}^2}}$$

Если tнабл>tкр=t(α,n - 2), где t - распределение Стьюдента, то отклоняется гипотеза об отсутствии гетероскедастичности. Если имеется несколько объясняющих переменных то проверка осуществляется для каждой отдельно.

2.2.1.3 Тест Парка

Строятся регрессии yi = Ь0 + blxi + ei. Определяются $\ln e_i^2$ =ln(y_i, - $\bar{y}$ i,)^2. Строится регрессия ln e_i^2 =α+ (βlnx_1 +v_i. Проверяется значимость β на основе t-статистики $t = \dfrac{\beta}{S_\beta}$. Если значим, то имеется гетероскедастичность.

Недостатками теста является зависимость от вида функции, и гетероскедастичности v_i.

2.2.1.4 Тест Глейзера

Похож на тест Парка. Используется уравнение регрессии

$$|e_i| = \alpha + \beta x_i^k + v_i$$

Изменяя k строят различные регрессии (k =...-1,-0.5,0.5, 1,...). Значимость β соответствует гетероскедастичности.

2.2.1.5 Тест Голдфелда-Квандта

Предполагается, что стандартное случайное отклонение $\sigma_i = \sigma(\varepsilon_i)$ пропорционально значению xi переменной X в i-ом наблюдении.

Упорядоченная по X выборка разбивается на 3 выборки размером k,n-2k,k.

Оцениваются регрессии для 1 и 3 выборки. Определяются $S_1 = \sum_{i=1}^{k} e_i^2$,

$$S_3 = \sum_{i=n-k+1}^{n} e_i^2$$

Строится статистика $F\frac{S_3}{S_1}$. Если Fнабл>Fкр=F(α, ν $_1$,ν_2), где α-уровень значимости, $\nu_1=\nu_2$=k-m-l, m - число объясняющих переменных, F- статистика Фишера, то имеется гетероскедастичность

Для множественной регрессии данный тест проводят для той объясняющей переменной которая в наибольшей степени связана с σ_i, или для всех объясняющих переменных. При этом должно быть k>m+1.

Если стандартное случайное отклонение $\sigma_i = \sigma(\varepsilon_i)$ обратно пропорционально значению xi переменной X в i-ом наблюдении, то m используется $F\frac{S_3}{S_1}$

2.2.2 Снижение гетероскедастичности.

2.2.2.1 Метод взвешенных наименьших квадратов. Используется при известных σ_i^2 для каждого наблюдения. Рассмотрим парную регрессию.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Значения каждой пары (xi,yi) делят на σ_i. Строится регрессия не имеющая гетероскедастичности.

$$y_i^* = \beta_0 z_i + \beta_i x_i^* + v_i \text{ , где } z_i = \frac{1}{\sigma_i}, y_i^* = \frac{y_i}{\sigma_i}, x_i^* = \frac{x_i}{\sigma_i}, v_i = \frac{\varepsilon_i}{\nu_i}$$

2.2.2.2 Неизвестные дисперсии отклонений.

1) предполагается что дисперсии пропорциональны xi $\sigma_i = \sigma^2 x_i$ (σ^2 - коэффициент пропорциональности), в этом случае уравнение регрессии преобразуется делением обеих частей на $\sqrt{x_i}$;

2) предполагается что дисперсии пропорциональны в этом случае, уравнение регрессии преобразуется делением обеих частей на xi.

2.3 Автокорреляция

Для обнаружения наиболее известным является критерий Дарбина-Уотсона. Имеет ограничения

1) применяется только для моделей имеющих свободный член;

2) случайные отклонения e_t определяются по авторегрессионной схеме 1-го порядка AR(1) $\varepsilon_t = \rho\varepsilon_{t-1} + v_t, v_t$ - случайный член;

3) не должно быть пропусков в наблюдениях;

4) в составе объясняющих переменных не должно быть зависимой переменной с временным лагом в один период, т.е. модель не должна иметь вид

$$y_t = \beta_0 + \beta_1 x_{t1} + ... + \beta_m x_{tm} + \gamma y_{t-1} + \varepsilon_t$$

Для устранения автокорреляции меняют спецификацию модели, вид зависимости. Если не помогает можно выполнить авторегрессионное преобразование. Для линейной регрессионной модели используется AR(1).

3 Реализация в математической системе

В качестве инструментальной системы выбрана система Mathcad. Которая является достаточно мощной и простой по реализации. Используется функция множественной полиномиальной регрессии – regress. Документ для её использования приведен ниже.

Число переменных:

$n = 12$

$z := \text{regress}(X, Y, k)$

$i := 0..N - 1$

Полином, соответствующий функции:

$\text{fit}(x) := \text{interp}(z, X, Y, x)$

$\text{predY}_i := \text{fit}\left[\left(X^T\right)^{\langle i \rangle}\right]$

Коэффициенты уравнения регрессии $y = a_0 + a_1x_1 + \ldots + a_nx_n$

$\text{coeffs} := \text{submatrix}(z, 3, \text{length}(z) - 1, 0, 0)$

Отклонение:

$\text{resid} := \text{predY} - Y$

Заключение

В области статистики постоянно возникают проблемы, с которыми сталкиваются наука и промышленность. Эти проблемы часто возникают в результате сельскохозяйственных и промышленных экспериментов. С появлением компьютеров и информационной эры статистические задачи резко возросли как по масштабам, так и по сложности. Проблемы в области хранения, организации и поиска данных привели к появлению новой области "интеллектуального анализа данных"; статистические и вычислительные задачи в биологии и медицине привели к появлению "биоинформатики". Огромные объемы данных обрабатываются данные генерируются во многих областях, и задача статистика – разобраться во всем этом: выделить важные закономерности и тенденции и понять, "о чем говорят данные". Используется понятие – обучение на основе данных.

Список использованных источников

1. С.Н. Лапач, А.В. Чубенко, П.Н. Бабич Статистические методы в медико-биологических исследованиях с использованием Excel. Экспериментальные исследования Клинические испытания Анализ фармацевтического рынка - К: МОРИОН, 2001.

2. Manish Rana, Dr.R.R.Sedamkar Design of expert system for medical diagnosis using fuzzy logic International Journal of Scientific & Engineering Research, Volume 4, Issue 6, June-2013 pp.2914-2921

3. Дубров А.М., Мхитарян В.С., Тропшн Л.И. Многомерные статистические методы. - М.: Финансы и статистика, 1988.