УДК 004.942

Уртаев Александр Казбекович
студент 2 курса магистратуры
факультет информационных технологий и электронной техники
Северо-Кавказский горно-металлургический институт
(государственный технологический университет),
Россия, г. Владикавказ
e-mail: urtaev_2012@mail.ru

ОЦЕНКА И ВЫБОР МАТЕМАТИЧЕСКОЙ МОДЕЛИ ДЛЯ ПРОГНОЗИРОВАНИЯ САХАРНОГО ДИАБЕТА

Аннотация: *В данной статье рассматривается краткая характеристика заболевания «сахарный диабет» (СД) для рассмотрения процесса построения, оценки и выбора математической модели множественной регрессии с помощью инструментария языка программирования Python. После оценки и выбора лучшей модели начинается работа по прогнозированию наблюдаемых процессов.*

Ключевые слова: Сахарный диабет, математическое моделирование, множественная регрессия, разработка, оценка, отбор, язык программирования, Python, прогнозирование.

Urtaev Alexander Kazbekovich
2nd year master student,
Faculty of Information Technology and Electronic Engineering
North Caucasian Institute of Mining and Metallurgy (State Technological University),
Russia, Vladikavkaz

ASSESSMENT AND SELECTION OF A MATHEMATICAL MODEL FOR PREDICTION OF DIABETES MELLITUS

Abstract: *This article discusses a brief description of the disease "diabetes mellitus" (DM) to consider the process of building, evaluating and choosing a mathematical model of multiple regression using the tools of the Python programming language. After evaluating and choosing the best model, work begins on predicting the observed processes.*

Key words: Diabetes mellitus, mathematical modeling, multiple regression, development, forecasting, selection, programming language, Python, estimation.

Введение

В медицинской практике не ново явление использования математических методов и решений в различных отраслях ее деятельности. Ученые на стыке таких наук как медицина, биология и математика посвящают много времени для интеграции математических моделей, так как специфика и проблематика того или иного заболевания имеют сложный характер. В настоящее время большое множество заболеваний имеют достаточную или полную методологию, что в разы облегчают получение данных наблюдений, на основе которых производится автоматизированное прогнозирование и имитации болезни.

СД является предметом исследований и разработок в течении десятилетий. Это заболевание преследует человечество с давних времен и только совсем недавно его научились правильно диагностировать и лечить. Как и любое другое заболевание, сахарный диабет исследуется с помощью современных методов и решений статистики и программирования.

Характеристика СД

СД – клинический синдром хронической гипергликемии и глюкозурии, обусловленный абсолютной или относительной инсулиновой недостаточностью, приводящей к нарушению обмена веществ, поражению сосудов (различные ангиопатии), нейропатии и патологическим изменениям в различных органах и тканях [1, с. 51].

При классификации СД выделяет 2 типа:

A. инсулинозависимый – I тип;

B. инсулиннезависимый – II тип [2, с. 219].

Первый тип характеризуется неспособностью организма к выработке инсулина. Второй тип обусловлен низким содержанием или низкой работоспособностью инсулина.

Для лечения СД нужно следовать строгим предписаниям лечащего врача. Так для лечения инсулинозависимого типа нужно получать инсулин извне по средствам специальный медицинских приспособлений виде шприц ручке или помпы. Для Лечения инсулиннезависимого типа лечащий врач назначает специальную диету и физические нагрузки. В случае если этого недостаточно

применяет лекарственные препараты, помогающие вырабатывать достаточное количество инсулина.

Обычно II тип СД при несвоевременном лечении переходит в I тип после чего жизнь больного подвергается большому риску. Для поддержания жизнедеятельности организма при I типе СД можно разработать математическую модель, по которой можно понять поведение сахара в крови, выявить углеводный коэффициент и спрогнозировать результат ввода инсулина.

Построение математической модели

В моделировании очень важно, чтобы получаемая модель была простой и понятной, а также она должна очень точно описывать наблюдаемые процессы. В нашем случае будем использовать метод наименьших квадратов (МНК) за основу модели.

МНК в настоящее время широко применяется при обработке количественных результатов естественно-научных опытов, технических данных, астрономических и геодезических наблюдений и измерений [3, с. 7].

Его суть заключается в минимизации суммы квадратов отклонения построенной функции от тех переменных, что нам известны. Преимущества МНК в простоте и применимости, однако она может быть ненадежна, когда статистика данных распределяется не как обычно. Но для многих точек эта проблема решаема.

Данные для моделирования были взяты с дневника одного из добровольца. Чтобы пользователь мог быстрее использовать модель берется небольшая выборка из 20 значений и примерная функция, которая больше походит на результаты зависимой переменной.

Моделирование в этом случае осуществляется по средствам инструментария языка программирования Python в среде разработки PyCharm. Python универсальных язык программирования. Используется в статистике, разработке программного обеспечения, мобильных приложений и создании сайтов. Отличный современный вариант, который насчитывает тысячи готовых библиотек.

Построим множественную регрессионную модель с двумя переменными M1 (1):

$$Y = B_0 + B_1 * X1 + B_2 * X1^2 + B_3 * X2 + B_4 * X1 * X2 \quad (1)$$

где Y – инсулин, X1 – порядковый номер измерения сахара в крови, X2 – показатель сахара.

Имеются 2 матрицы: 1 матрица независимых переменных, 2 матрица зависимых переменных. Получение коэффициентов регрессии осуществляется по формуле (2):

$$B = (X^T * X)^{-1} * X^T * Y \quad (2)$$

Реализация на языке Python выглядит продемонстрирована на рисунке 1:

```
37    b1 = X.transpose()
38    b2 = b1.dot(X)
39    b3 = np.linalg.inv(b2)
40    b4 = b3.dot(b1)
41    B = b4.dot(Y)
```

Рисунок 1 – коэффициенты регрессии.

Где: X – матрица независимых переменных;

Y – матрица фактических значений зависимых переменных.

Таблица коэффициентов B:

Таблица 1.

Коэффициенты регрессии

B_0	4.87729206e+00
B_1	-1.11078832e-01
B_2	1.64537430e-03
B_3	1.66860124e-01
B_4	8.34441087e-03

Для признания адекватности модели нужно найти F критерий Фишера расчетный (Fr) и сравнить с табличным значением (F). Расчетный критерий Фишера есть отношение между дисперсией зависимой переменной и дисперсией адекватности.

Дисперсия адекватности рассчитывается как отношение суммы квадрата разности Y фактических с Y расчетных (Yr) значений и разности количества измерений (N) с количеством коэффициентов (k) регрессии (3):

$$\mathrm{Dad} = \frac{\sum(\mathrm{Y} - \mathrm{Yr})^2}{\mathrm{N} - \mathrm{k}} \tag{3}$$

Для нашей модели она составляет - 0.03750403.

Дисперсия зависимой переменной рассчитывается как отношение суммы квадрата разности Y фактических с Y средних значений (Ysr) и разности количества измерений (N) и единицы (4):

$$\mathrm{DY} = \frac{\sum(\mathrm{Y} - \mathrm{Ysr})^2}{\mathrm{N} - 1} \tag{4}$$

Для нашей модели она составляет – 0.33418421.

Расчетное значение F – статистики – 8.9106213. Мы берем уровень значимости 0,01, так как от качества модели зависит состояние пациента, и мы должны добиться лучшей точности.

Табличное значение F – критерия Фишера – 2.339819281665458.

Вывод: уравнение регрессии признано адекватным экспериментальным данным на уровне значимости 0,01, что соответствует доверительной вероятности p = 99,0%, т.к. Fr > R.

Построим доверительные интервалы для прогнозирования зависимой переменной на рисунке 2:

```
G = b3
d0 = t * math.sqrt(Dad * G[0, 0])
d1 = t * math.sqrt(Dad * G[1, 1])
d2 = t * math.sqrt(Dad * G[2, 2])
d3 = t * math.sqrt(Dad * G[3, 3])
d4 = t * math.sqrt(Dad * G[4, 4])
print('Доверительные интервалы коэффициентов регрессии ')
print(d0, d1, d2, d3, d4, sep='\n')
```

Рисунок 2 – Доверительные интервалы.

Сравнивая последовательно коэффициенты регрессии B_0, B_1, B_2, B_3, B_4 с d0, d1, d2, d3, d4 соответственно мы можем исключить их так как в таком случае они являются не значимыми, но только если они меньше соответствующим значениям доверительных интервалов для коэффициентов модели.

Для прогнозирования зависимой переменной мы переходим на шаг вперед подставляя независимые переменные и по найденному уравнению получаем значения зависимой переменной. На рисунке 3 продемонстрирован код, который также составляет график функции (рисунок 4):

```
t1 = 21
A = 9

YP = B[0] + B[1] * t1 + B[2] * (t1 ** 2) + B[3] * A + B[4] * t1 * A
print('YP =', YP, '\n')
Ymax = YP + S[-1]
YRmax = [x+y for x, y in zip(YR, S)]
print('Ymax =', Ymax, '\n')
print('YRmax =', YRmax, '\n')
Ymin = YP - S[-1]
YRmin = [x-y for x, y in zip(YR, S)]
print('Ymin =', Ymin, '\n')
print('YRmin =', YRmin, '\n')

x1 = np.arange(0, N).reshape(-1, 1)
print('x1 =', x1, sep='\n')
plt.plot(x1, YR, label='YR')
plt.plot(x1, Y, label='Y', marker='o')
plt.plot(x1, YRmin, label='YRmin')
plt.plot(x1, YRmax, label='YRmax')
plt.legend(loc='best')
plt.grid()
plt.show()
```

Рисунок 3 – Прогнозирование.

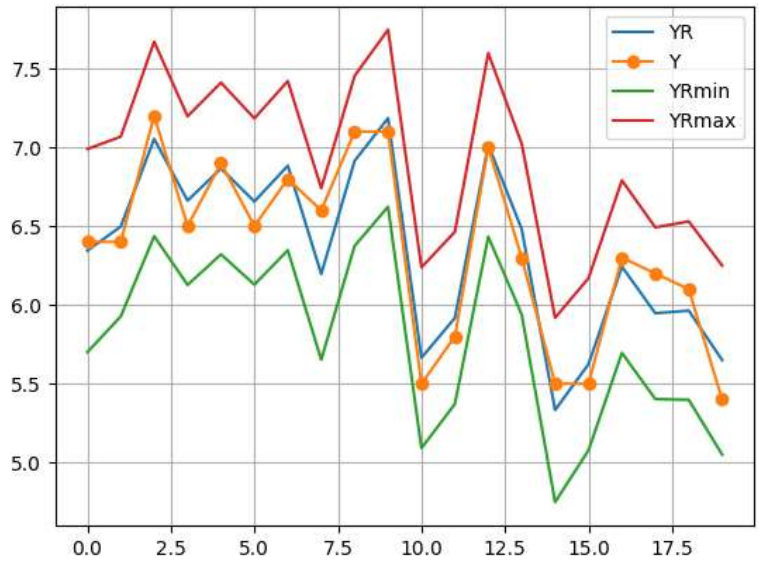

Рисунок 4 – График функции.

Прежде чем использовать модель для прогнозирования нужно проверить, лучше всего ли описывает данная модель наблюдаемый нами процесс. То есть нам необходимо провести сравнение показателей этой модели с различными вариациями или других типов. Для примера составим 2 вариации данной модели, продемонстрированные на формулах М2 (5) и М3 (6):

$$Y = B_0 + B_1 * X1 + B_2 * X1^2 + B_3 * X2 + B_4 * X2^2 + B_5 * X1*X2 \quad (5)$$

$$Y = B_0 + B_1 * X1 + B_2 * X2 + B_3 * X2^2 + B_4 * X1*X2 \quad (6)$$

Построим таблицу основных показателей и коэффициентов для наглядной оценки и выбора лучшей модели.

Таблица 2.

Основные показатели моделей

	M1	**M2**	**M3**
corr	0,9547	0,9547	0,6524
corr кр.	4,3487	4,3487	4,3482
Fr	8,9106	8,3182	8,5018
F	2,3398	2,4001	2,3398
Fr/F	3,8083	3,4659	3,6335
R^2	0,9114	0,9114	0,9071
R^2adj	0,8788	0,8705	0,8719

MSE	0,02813	0,02812	0,0295
RMSE	0,1677	0,1677	0,1717
Абсолютная ошибка	0,0451	0,0504	0,0745
Относительная ошибка	0,7791	0,8002	1,1818

Где:

corr – коэффициент корреляции расчетных значений зависимой переменной с фактическими;

corr – критический коэффициент корреляции;

Fr – Расчетное значение критерия Фишера;

F – табличное значение критерия Фишера;

R^2 – коэффициент детерминации;

R^2adj – скорректированный коэффициент детерминации;

MSE – средняя квадратичная ошибка;

RMSE – корень из средней квадратичной ошибки;

Абсолютная ошибка – модуль разницы Y фактического от Y прогнозного;

Относительная ошибка – отношение Абсолютной ошибки на Y фактического, считается в процентах.

Сравнивая основные показатели моделей можно сделать вывод, что модель M1 лучше подходит для описания наблюдаемого процесса так как имеет наибольшее Fr и наименьшее абсолютную и относительную ошибки.

Заключение

Это, конечно, не весь процесс оценки и выбора лучшей модели. Главное было показать, что какими методами нужно руководствоваться при разработке и отбора нужного инструментария для проведения исследований в медицинских или других отраслях науки. Для математических моделей множественной регрессии важно адекватно и точно описывать наблюдаемые процессы. На выбор влияет как сложность модели, так и средние квадратичные, корень средних квадратических, абсолютная и относительная ошибки.

Список литературы:

1. Балаболкин М.И. Эндокринология. М.: «Универсум Паблишинг», 1998. 580 с.

2. Старкова Н.Т. Клиническая эндокринология: Руководство. М.: Медицина, 1991. 576 с.

3. Линник Ю. В. Метод наименьших квадратов и основы математико-статистической теории обработки наблюдений. М.: Государственное издательство физико-математической литературы, 1962. 352 с.