УДК 004.42:519.24

МОДЕЛЬ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ ДЛЯ ДИАГНОСТИКИ СЕРДЕЧНО-СОСУДИСТЫХ ЗАБОЛЕВАНИЙ

Асанова А.Э., магистрант гр. ПИм-191, II курс
Заболотин А.А., магистрант гр. ПИм-191, II курс
Научный руководитель: Пимонов А.Г., д.т.н., профессор
Кузбасский государственный технический университет
имени Т.Ф. Горбачева
г. Кемерово

По данным Всемирной организации здравоохранения ежегодно от болезней сердечно-сосудистых заболеваний (ССЗ) умирает 17,5 млн человек по всему миру. На 2016 год от этих болезней умерло порядка 17,9 млн человек, что составляет 31 % от всех случаев смерти в мире. При этом 85 % [1] смертей являются результатом сердечного приступа или инсульта.

Зависимость одной случайной величины от значений, которые принимает другая случайная величина (физическая характеристика), в статистике называется корреляцией. Если этой зависимости придан аналитический вид, то такую форму представления изображают уравнением регрессии. Процедура поиска вида предполагаемой зависимости между различными числовыми совокупностями обычно включает следующие этапы:

- установление значимости связи между ними;
- возможность представления этой зависимости в форме математического выражения (уравнения регрессии).

Первый этап в указанном статистическом анализе касается выявления так называемой корреляции, или корреляционной зависимости. Корреляция рассматривается как признак, указывающий на взаимосвязь ряда числовых последовательностей. Иначе говоря, корреляция характеризует силу взаимосвязи [2, 3] в данных. Второй этап – это регрессионный анализ. Это один из методов исследования свойств данных и их моделирования. Параметры регрессионной модели настраиваются таким образом, чтобы на основе независимых переменных выходные (модельные) данные были максимально приближены к известным зависимым.

Исходные данные для анализа были взяты на платформе Kaggle [4]. Это публичная веб-платформа, в которой размещены наборы открытых данных. Выборка содержит 70 тыс. записей, для анализа были использованы 10 тыс. наблюдений. Переменные, выбранные для анализа, приведены в таблице 1.

Таблица 1 – Переменные для анализа, измеренные в разных шкалах

Интервальная	**Ранговая**		**Номинальная**	
Название	Название	Пояснения	Название	Пояснения
Возраст	Холестерин	1 -Норма 2 -Выше нормы 3 -Сильное превышение	Курение	0 -Да 1 -Нет
Рост	Глюкоза	1 -Норма 2 -Выше нормы 3 -Сильное превышение	Пол	1 -Жен. 2 -Муж.
Вес			Алкоголь	0 -Да 1 -Нет
Систолическое давление			Физ. активность	0 -Да 1 -Нет
Диастолическое давление			Болезнь	0 -Да 1 -Нет

Переменная «*Болезнь*» является выходной переменной. В данном случае ставится задача классификации. В качестве объекта выступает пациент, которого необходимо отнести к одному из классов: больной, здоровый. Влияющими признаками считаются результаты обследования, симптомы, общие сведения.

В выборке присутствуют данные, относящиеся к ранговой шкале («*Холестерин*», «*Глюкоза*»), поэтому в работе применялся подсчет коэффициента корреляции рангов Спирмена. Так как прогнозируемый фактор «*Болезнь*» представляет собой бинарную переменную [5], то использовалась логистическая регрессия. Это разновидность множественной регрессии, в которой используется логистическая функция, что позволяет смоделировать зависимость бинарной выходной переменной от набора входных данных (предикторов). На выходе производится не предсказание числовой переменной, а ее вероятность принадлежности к одному из классов.

В работе проведен сравнительный анализ двух регрессионных моделей, полученных разными методами. Выбор метода отбора [5] позволяет задать то, каким образом независимые переменные включаются в модель. Используя различные методы, можно построить целый ряд регрессионных моделей для одного и того же набора переменных. Первый метод «*Enter*», когда все переменные включаются в модель на первом шаге. И второй – «*Обратное исключение* (*Вальд*)» или шаговый отбор исключением. Проверка на исключение основана на использовании статистики Вальда в качестве критической.

Для оценки прогностической силы моделей воспользуемся ROC-анализом. Площадь под ROC-кривой является меркой качества для модели бинарной классификации. В случае идеальной модели график кривой ROC проходит через точку (0,1), а площадь под ней максимальна и равна 1. Для построения такого графика необходимо подсчитать таблицы классификации. Это измерение производительности для задачи классификации машинного обучения, когда на выходе может быть два или более классов. В результате получается таблица с 4 различными комбинациями расчетных и фактических значений: истинно-положительными, ложно-положительными, истинно-отрицательными и ложно-отрицательными.

Исследование проводилось средствами MS Excel и SPSS Statistics. Последняя – одна из самых распространенных программ для обработки статистической информации. Она включает в себя большое количество модулей и позволяет обрабатывать большие объемы данных.

Перед работой была проведена проверка на анормальность данных, в ходе которой были исключены 298 записей. Ранговые параметры («*Холестерин*», «*Глюкоза*») были разбиты на отдельные бинарные столбцы.

Матрица корреляции Спирмена приведена на рисунке 1.

B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
	Кардио	Возраст	Пол	Рост	Вес	Давление Систол,	Давление Дистол,	Курение	Алкоголь	Физ,активность	Холестерин=1.0	Холестерин=2.0	Холестерин=3.0	Глюкоза=1.0	Глюкоза=2.0	Глюкоза=3.0
Кардио	1,00															
Возраст	-0,24	1,00														
Пол	0,00	0,01	1,00													
Рост	-0,02	0,09	0,54	1,00												
Вес	0,18	-0,05	0,17	0,31	1,00											
Давление Систол	**0,45**	-0,23	0,05	0,02	0,26	1,00										
Давление Дистол	**0,35**	-0,16	0,06	0,03	0,24	0,73	1,00									
Курение	-0,02	0,05	0,33	0,20	0,07	0,01	0,01	1,00								
Алкоголь	0,00	0,04	0,18	0,12	0,09	0,03	0,04	0,33	1,00							
Физ,активность	-0,04	0,02	0,02	0,00	-0,02	0,00	0,00	0,03	0,04	1,00						
Холестерин=1.0	-0,20	0,14	0,04	0,06	-0,12	-0,20	-0,16	-0,03	-0,03	-0,01	1,00					
Холестерин=2.0	0,06	-0,03	-0,03	-0,04	0,06	0,09	0,08	0,03	0,03	-0,01	-0,69	1,00				
Холестерин=3.0	0,20	-0,17	-0,02	-0,04	0,10	0,18	0,14	0,01	0,01	0,03	-0,62	-0,14	1,00			
Глюкоза=1.0	-0,08	0,08	0,02	0,03	-0,11	-0,10	-0,07	-0,03	-0,03	0,00	0,36	-0,16	-0,33	1,00		
Глюкоза=2.0	0,04	-0,02	0,00	-0,02	0,09	0,06	0,04	0,04	0,05	-0,02	-0,20	0,27	-0,02	-0,68	1,00	
Глюкоза=3.0	0,07	-0,08	-0,03	-0,03	0,06	0,07	0,05	-0,01	-0,01	0,01	-0,30	-0,06	0,47	-0,67	-0,08	1,00

Рисунок 1 – Матрица корреляции Спирмена

Средняя степень влияния на зависимую переменную» наблюдалась только у диастолического (0,35) и систолического (0,45) давлений. А факторы «*Алкоголь*» и «*Пол*» имеют коэффициенты корреляции меньше 0,1. Т. е. данные параметры не оказывают влияния на наличие болезни, и возможно их не включать в окончательную модель.

Далее были построены регрессионные модели и таблицы классификации для каждого из выбранных методов (рис.2, 3).

Переменные в уравнении.(Enter)

	Возраст	Рост	Вес	Давление Систол	Давление Дистол	Курение	Физактивность	Холестерин_1	Холестерин_2	Глюкоза_1	Глюкоза_2	Пол	Алкоголь	Константа
Шаг 1	-,049	-,006	,014	,053	,018	-,121	-,250	-1,112	-,856	,297	,292	-,034	-,112	-3,860

Переменные в уравнении.(Назад: Вальда)

	Возраст	Рост	Вес	Давление Систол	Давление Дистол	Курение	Физактивность	Холестерин_1	Холестерин_2	Глюкоза_1	Глюкоза_2	Константа
Шаг 3	-,049	-,007	,014	,053	,018	-,164	-,252	-1,111	-,855	,294	,286	-3,700

Рисунок 2 – Регрессионные модели

Таблица классификации (Enter)

Наблюденные		Предсказанные		
		Кардио		Процент правильных
		Здоров	Болен	
Кардио	Здоров	3861	1014	79,2
	Болен	1626	3201	66,3
Общая процентная доля				72,8

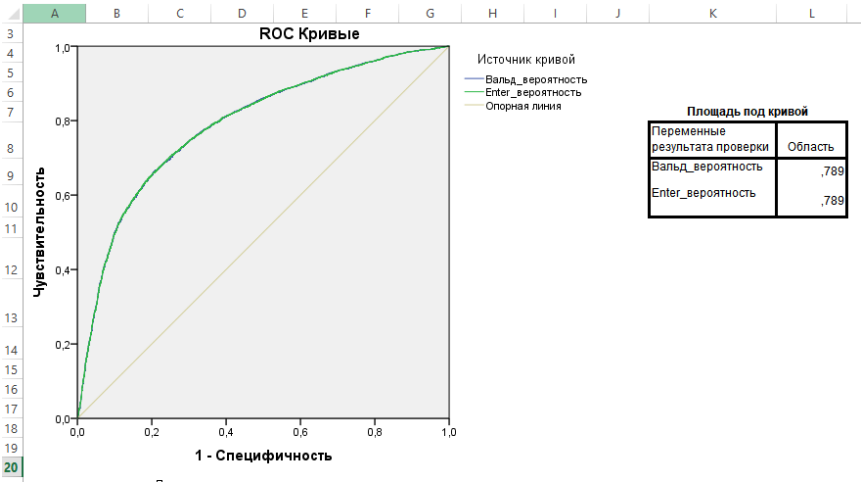
Таблица классификации (Назад: Вальда)

Наблюденные			Предсказанные		
			Кардио		Процент правильных
			Здоров	Болен	
Шаг 1	Кардио	Здоров	3861	1014	79,2
		Болен	1626	3201	66,3
	Общая процентная доля				72,8
Шаг 2	Кардио	Здоров	3866	1009	79,3
		Болен	1627	3200	66,3
	Общая процентная доля				72,8
Шаг 3	Кардио	Здоров	3867	1008	79,3
		Болен	1625	3202	66,3
	Общая процентная доля				72,9

Рисунок 3 – Таблицы классификации

По таблицам классификации можно определить, что точность модели около 73 % для обоих методов. Также в модель полученным шаговым отсечением включены не все переменные, а именно отсутствуют «*Пол*» и «*Алкоголь*». Учитывая примерно одинаковую точность предсказания в обоих случаях, можно сделать вывод о их незначимости.

Оценка качества полученных регрессионных моделей с помощью ROC-анализа приведена на рисунке 4.



Площадь под кривой

Переменные результата проверки	Область
Вальд_вероятность	,789
Enter_вероятность	,789

Рисунок 4 – Результаты ROC-анализа

С вероятностью 79 % моделям удается разделить классы верно. По таблицам классификации, представленным на рис. 3, можно вычислить процент ошибок. Так вероятность ошибки при определении здорового человека равна 25 %, а больного 34 %. Ошибки прогноза достаточно велики, и приведенная модель требует дальнейшего уточнения.

Список литературы:

1. Федеральная служба государственной статистики [Электронный ресурс]. – Режим доступа: http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/publications/catalog/doc_1139919134734, свободный (дата обращения: 25.09.2020)

2. Бараз, В.Р. Корреляционно-регрессионный анализ связи показателей коммерческой деятельности с использованием программы Excel: учебное пособие. – Екатеринбург : ГОУ ВПО «УГТУ–УПИ», 2005. – 102 с.

3. Дороганов, В.С. Методы статистического анализа и нейросетевые технологии для прогнозирования показателей качества металлургического кокса / В.С. Дороганов, А.Г. Пимонов // Вестник Кемеровского государственного университета. – 2014. – № 4, Т. 3. – С. 123-129.

4. Cardiovascular Disease dataset // Kaggle [Электронный ресурс]. – Режим доступа: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset, свободный (дата обращения: 29.03.2021).

5. Методы отбора переменных для логистической регрессии // IBM Documentation [Электронный ресурс]. – Режим доступа: https://www.ibm.com/docs/ru/spss-statistics/25.0.0?topic=SSLVMB_25.0.0/spss/regression/logistic_regression_methods.html, свободный (дата обращения: 29.03.2021).