Федеральное государственное бюджетное учреждение
«Национальный медицинский исследовательский центр онкологии
имени Н.Н. Петрова» Министерства здравоохранения Российской Федерации
(ФГБУ «НМИЦ онкологии им. Н.Н. Петрова» Минздрава России)
Отдел учебно-методической работы

Федеральное государственное бюджетное образовательное учреждение
высшего образования «Северо-Западный государственный
медицинский университет имени И.И. Мечникова»
Министерства здравоохранения Российской Федерации
(ФГБОУ ВО СЗГМУ им. И.И. Мечникова Минздрава России)
Кафедра онкологии

Беляев А. М., Михнин А. Е., Рогачев М. В.

ROC-анализ и логистическая регрессия в MedCalc

Учебное пособие

Санкт-Петербург
2023

УДК:614.2:303.71(07)
ББК:51.1(2)я7

Беляев А. М., Михнин А. Е., Рогачев М. В. ROC-анализ и логистическая регрессия в MedCalc: учебное пособие для врачей и обучающихся в системе высшего и дополнительного профессионального образования. – Санкт-Петербург: НМИЦ онкологии им. Н.Н. Петрова, 2023. – 36 с.

ISBN 978-5-6048249-5-5

Рецензент: доктор медицинских наук, профессор В. М. Мерабишвили, председатель научно-медицинского Совета по развитию информационных систем онкологической службы СЗФО Российской Федерации, заведующий научной лабораторией онкологической статистики федерального государственного бюджетного учреждения «Национальный медицинский исследовательский центр онкологии имени Н.Н. Петрова» Министерства здравоохранения Российской Федерации, г. Санкт-Петербург.

В учебном пособии представлены доступное для неподготовленного читателя объяснение принципов, лежащих в основе логистической регрессии, и пошаговые инструкции по созданию логистических моделей в популярном статистическом пакете MedCalc. Аналогичным образом логистическое модели создаются и в более продвинутых статистических пакетах. Выполнение практических рекомендаций, изложенных в пособии, позволит избежать типичных ошибок, встречающихся в публикациях и диссертационных работах.

Учебное пособие предназначено для аспирантов, клинических ординаторов и врачей, начинающих свою научную деятельность в медицине, а также для обучающихся по программам дополнительного профессионального образования.

Утверждено в качестве учебного пособия
Ученым советом ФГБУ «НМИЦ онкологии
им. Н.Н. Петрова» Минздрава России
протокол № 4 от 18 апреля 2023 г.

ISBN 978-5-6048249-5-5

Содержание

Введение 4

Глава 1. ROC-анализ и логистическая регрессия в MedCalc 5

Глава 2. Логистическая регрессия 16

2.1. Определения 16

2.2. Математическое обоснование 18

2.3. Взаимосвязь логистической функции и вероятности 19

2.4. Условия применения и выбор предикторов 20

Глава 3. Логистическая регрессия в MedCalc 22

3.1. Оценка качества логистической регрессионной модели в MedCalc 25

3.2. Интерпретация уравнения логистической регрессии 31

Контрольные вопросы 33

Тестовые задания 33

Список литературы 36

Введение

В биомедицинских исследованиях нередко возникает задача разделения объектов на два класса, например, на здоровых и больных, на имеющих хороший и плохой прогноз заболевания и т.д. Наиболее удобным инструментом для решения задачи бинарной классификации является метод логистической регрессии, реализованный в большинстве статистических пакетов. В его основе лежит ROC-анализ, разработанный первоначально для военных технологий, а позднее нашедший широкое применение в экономике, финансах и биомедицине.

Классический дискриминантный анализ является параметрическим, требует нормальности распределений предикторов и малопригоден для небольших выборок. Для построения дискриминантной функции применяется метод наименьших квадратов. Логистическая регрессия основана на непараметрических критериях, не предъявляет специальных требований к форме распределения переменных и пригодна для небольших выборок. Для построения дискриминантной функции используется метод максимального правдоподобия.

Метод логистической регрессии удобно реализован в статистическом пакете MedCalc, который мы и рекомендуем начинающим исследователям. Вместе с тем, в более мощных статистических пакетах (SPSS, NCSS, Statistica) этот метод осуществляется аналогичным образом.

Глава 1.
ROC-анализ и логистическая регрессия в MedCalc

Определения

ROC-receiver operation characteristic – рабочая характеристика приемника: в радиолокации – способность приёмника выделять сигнал из помех.

ROC-анализ – метод наглядного сравнения и оценки качества моделей бинарных классификаторов с нахождением оптимального порога разделения для отнесения объектов к тому или иному классу путем построения ROC-кривых.

ROC-кривая – график, отображающий соотношение между чувствительностью алгоритма классификации и частотой ложноположительных ответов (значением 1-специфичность алгоритма) при пошаговом изменении порога решающего правила.

Из двух классов один называется классом с положительными исходами, второй – с отрицательными. ROC-кривая показывает зависимость количества верно классифицированных положительных примеров от количества неверно классифицированных отрицательных примеров.

В терминологии ROC-анализа первые называются истинно положительным, вторые – ложно отрицательным множеством. Классификатор имеет определенный параметр, варьируя который, можно проводить то или иное разбиение на два класса. Этот параметр называют порогом, или точкой отсечения (cut-off

value). В зависимости от него будут получаться различные величины ошибок I и II рода.

Чувствительность – способность алгоритма выявлять объекты, имеющие определенный признак, среди объектов, действительно имеющих данный признак. Для диагностического теста это способность выявлять больных среди действительно больных.

Специфичность – способность алгоритма выявлять объекты, не имеющие данного признака среди объектов действительно не имеющих данного признака. Для диагностического теста это способность выявлять здоровых среди действительно здоровых.

Задача алгоритма классификации состоит в том, чтобы относить ранее неизвестные объекты к тому или иному классу. Примером такой задачи может быть диагностика: пациент болен (положительный результат) или пациент здоров (отрицательный результат). Результатом классификации могут быть четыре варианта:

- истинно-положительный результат (true-positive, TP) – пациент больной, диагноз положительный;
- ложноположительный результат (false-positive, FP) – пациент здоров, диагноз положительный;
- истинно-отрицательный результат (true-negative, TN) – пациент здоров, диагноз отрицательный;
- ложноотрицательный результат (false-negative, FN) – пациент больной, диагноз отрицательный.

Результаты классификации могут быть представлены в виде таблицы сопряжённости размера 2 × 2 (табл. 1).

Таблица 1

Результаты классификации
[оригинальная таблица]

TEST	DISEASE	
	Present	Absent
Positive	*TP*	*FP*
Negative	*FN*	*TN*

Чувствительность теста: $\boldsymbol{Sen} = \frac{TP}{TP+FN}$

Специфичность теста: $\boldsymbol{Spe} = \frac{TN}{TN+FP}$

Частота ложноположительных ответов (1 – Spe), необходимая для построения ROC-кривой, может быть определена как:

$$1\text{-Spe} = \left(1 - \frac{TN}{TN+FP}\right) = \frac{TN+FP}{TN+FP} - \frac{TN}{TN+FP}, \text{ или}$$

$$\boldsymbol{1\text{-}Spe} = \frac{FP}{TN+FP}$$

Точность теста – доля всех правильных результатов:

$$\boldsymbol{Acc} = \frac{TP+TN}{TP+TN+FP+FN} \quad .$$

Диагностическое отношение шансов:

$$\boldsymbol{DOR} = \frac{\frac{TP}{FP}}{\frac{FN}{TN}} = \frac{TP*TN}{FP*FN} \ ,$$

т.е. отношение шансов теста быть положительным, если у субъекта есть заболевание, к шансам теста быть положительным, если у субъекта нет заболевания.

Индекс Юдена J (Youden index): $\boldsymbol{J}$ = Se + Sp-1.

Индекс Юдена – это разница между долей истинно положительных результатов (чувствительностью теста) и долей ложноположительных результатов. Чем больше это различие, тем лучше работает диагностический алгоритм.

Путь к модулю ROC curve analysis в MedCalc показан на рисунке 1.

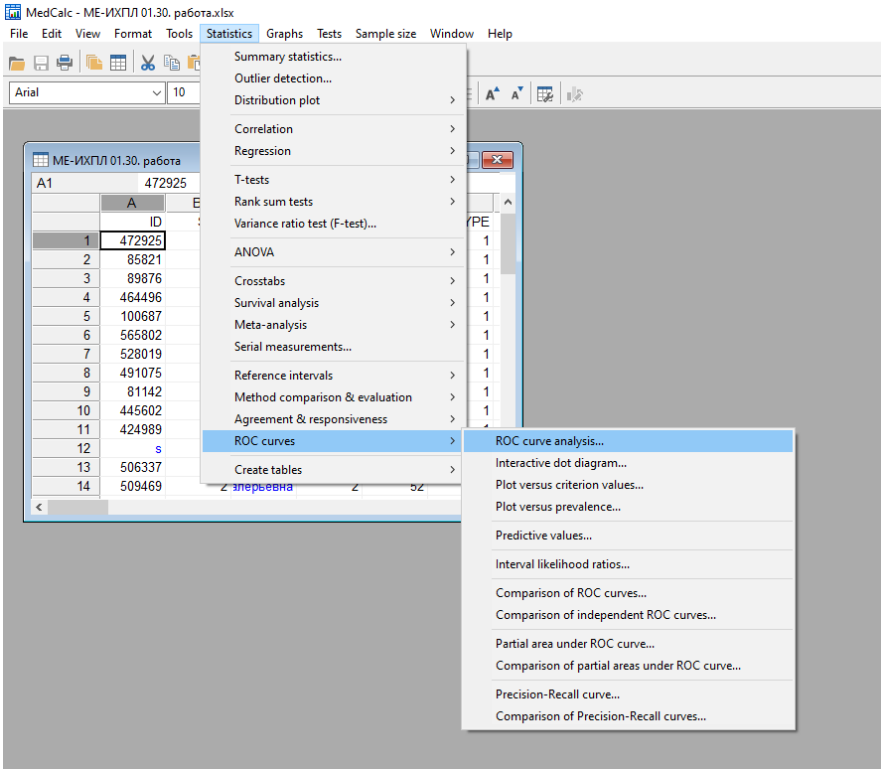


Рис. 1. Путь к модулю ROC curve analysis в MedCalc [оригинальный рисунок].

Выбор зависимой (определяющей класс объекта) и классифицируемой переменной (classification variable) для построения ROC кривой показан на рисунке 2.

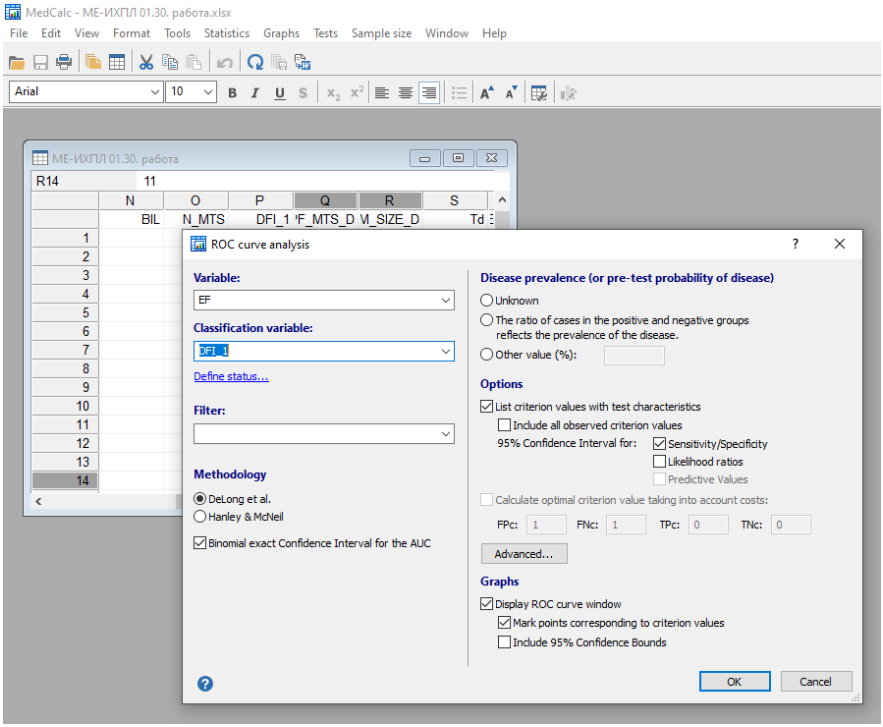


Рис. 2. Выбор зависимой (определяющей класс объекта) и классифицируемой переменной (classification variable) для построения ROC кривой [оригинальный рисунок].

Пример ROC-кривой приведен на рисунке 3.

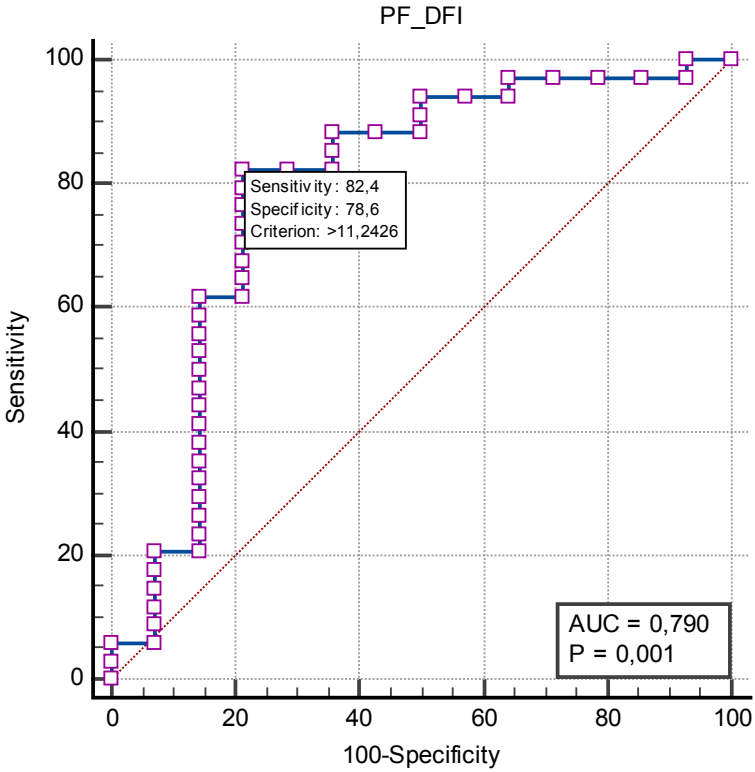


Рис. 3. ROC-кривая классификатора с указанием чувствительности, специфичности, AUC и уровня статистической значимости модели, построенная в MedCalc [оригинальный рисунок].

ROC-кривая строится следующим образом: для каждого значения порога отсечения, которое меняется от 0 до 1 с некоторым шагом, программа рассчитывает значения чувствительности Se и специфичности Sp. Строится график зависимости: по

оси Y фиксируется чувствительность Se (%), по оси X отмечается значение (100% – Sp) (сто процентов минус специфичность).

График часто дополняют диагональю y = x. Для идеального классификатора график ROC-кривой проходит через верхний левый угол, где доля истинно положительных случаев составляет 100% (идеальная чувствительность), а доля ложноположительных примеров равна нулю. Поэтому чем ближе кривая к верхнему левому углу, тем выше предсказательная способность модели. Наоборот, чем меньше изгиб кривой и чем ближе она расположена к диагонали, тем модель менее эффективна. Диагональная линия соответствует «бесполезному» классификатору, который сортирует объекты случайным образом.

При построении ROC-кривой MedCalc формирует таблицу характеристик классификатора, позволяющих оценить его качество. Из нескольких классификаторов лучшим считают классификатор, обеспечивающий наибольшую площадь под ROC-кривой (AUC) при высоком уровне значимости.

Программа перечисляет чувствительность и специфичность алгоритма классификации по каждой из точек отсечения. Оптимальный критерий разделения на классы отмечается на графике (рис. 3, табл. 2, criterion > 11,2426).

Таблица 2

Характеристики классификатора (теста), сформированные модулем построения ROC кривой в MedCalc [оригинальная таблица]

ROC curve

Variable	PF_DFI
Classification variable	EF

Sample size	48
Positive group [a]	34 (70,83%)
Negative group [b]	14 (29,17%)

[a] EF = 1
[b] EF = 0

Disease prevalence (%)	unknown

Area under the ROC curve (AUC)

Area under the ROC curve (AUC)	**0,790**
Standard Error [a]	0,0856
95% Confidence interval [b]	0,648 to 0,894
z statistic	3,387
Significance level P (Area=0.5)	0,0007

[a] DeLong et al., 1988
[b] Binomial exact

Youden index

Youden index J	0,6092
Associated criterion	>11,24260355
Sensitivity	**82,35**
Specificity	**78,57**

Criterion values and coordinates of the ROC curve [Hide]

Criterion	Sensitivity	95% CI	Specificity	95% CI	+LR	-LR
≥2,235371466	100,00	89,7 - 100,0	0,00	0,0 - 23,2	1,00	
>2,235371466	100,00	89,7 - 100,0	7,14	0,2 - 33,9	1,08	0,00
>3,879026956	97,06	84,7 - 99,9	7,14	0,2 - 33,9	1,05	0,41
>5,621301775	97,06	84,7 - 99,9	35,71	12,8 - 64,9	1,51	0,082
>5,79	94,12	80,3 - 99,3	35,71	12,8 - 64,9	1,46	0,16
>7,462195924	94,12	80,3 - 99,3	50,00	23,0 - 77,0	1,88	0,12
>8,744247206	88,24	72,5 - 96,7	50,00	23,0 - 77,0	1,76	0,24
>9,007232084	88,24	72,5 - 96,7	64,29	35,1 - 87,2	2,47	0,18
>9,500328731	82,35	65,5 - 93,2	64,29	35,1 - 87,2	2,31	0,27
>11,24260355	**82,35**	**65,5 - 93,2**	**78,57**	**49,2 - 95,3**	**3,84**	**0,22**
>17,061143984	61,76	43,6 - 77,8	78,57	49,2 - 95,3	2,88	0,49
>17,324128863	61,76	43,6 - 77,8	85,71	57,2 - 98,2	4,32	0,45
>39,776462853	20,59	8,7 - 37,9	85,71	57,2 - 98,2	1,44	0,93
>42,932281394	20,59	8,7 - 37,9	92,86	66,1 - 99,8	2,88	0,86
>82,807363577	5,88	0,7 - 19,7	92,86	66,1 - 99,8	0,82	1,01
>86,028928337	5,88	0,7 - 19,7	100,00	76,8 - 100,0		0,94
>99,342537804	0,00	0,0 - 10,3	100,00	76,8 - 100,0		1,00

В MedCalc есть возможность наглядного сравнения ROC-кривых (рис. 4 и 5).

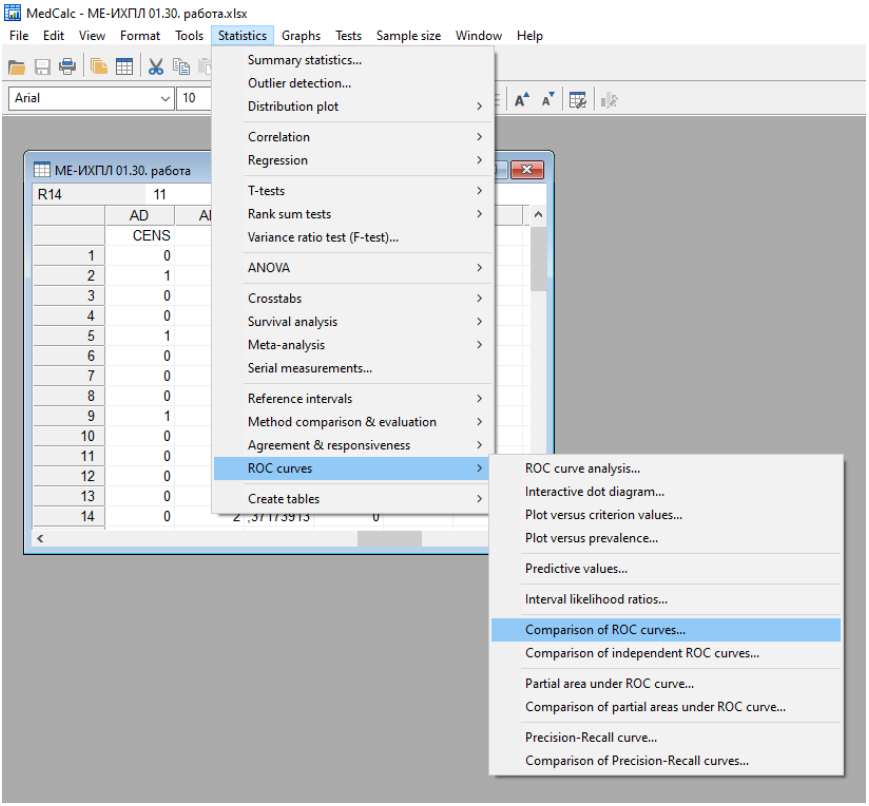
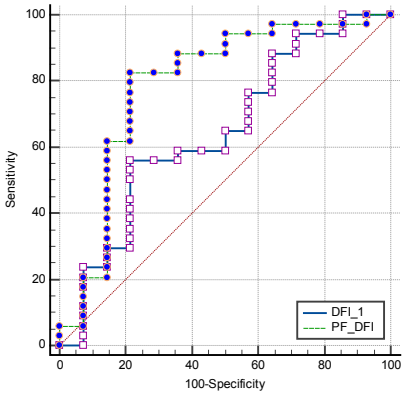


Рис. 4. Путь к модулю сравнения ROC-кривых [оригинальный рисунок].



Comparison of ROC curves

Variable 1	DFI_1
Variable 2	PF_DFI
Classification variable	EF

Sample size	48
Positive group [a]	34 (70,83%)
Negative group [b]	14 (29,17%)

[a] EF = 1
[b] EF = 0

Variable	AUC	SE [a]	95% CI [b]
DFI_1	0,643	0,0914	0,492 to 0,776
PF_DFI	0,790	0,0834	0,648 to 0,894

[a] Hanley & McNeil, 1982
[b] Binomial exact

Pairwise comparison of ROC curves

DFI_1 ~ PF_DFI	
Difference between areas	0,147
Standard Error [a]	0,0694
95% Confidence Interval	0,0111 to 0,283
z statistic	2,120
Significance level	**P = 0,0340**

Рис. 5. Сравнение двух ROC-кривых в MedCalc [оригинальный рисунок].

Pairwise comparison of ROC curves. Парное сравнение ROC-кривых на рис. 5. Разница AUC = 0,147, уровень статистической значимости p=0,034.

Глава 2.
Логистическая регрессия

2.1. Определения

Регрессия **–** способ предсказания значения одних переменных по значениям других.

Регрессионная модель – уравнение, в котором зависимая переменная представлена в виде функции от независимых переменных (регрессоров, предикторов).

Отношение шансов (Odds Ratio, *OR*) **–** отношение вероятности того, что событие произойдет *p(X)* к вероятности того, что событие не произойдет *1-p(X):*

$$OR = \frac{\boldsymbol{p(X)}}{\boldsymbol{1-p(X)}}$$

Логит – натуральный логарифм отношения шансов:

$$\text{logit (p)} = \ln(OR)$$

Нередко в исследованиях возникает вопрос оценки частот положительного и отрицательного исходов (бинарный отклик) в зависимости от факторов, представленных как дискретными, так и непрерывными переменными. Для такой оценки применяется логистическая регрессия. В качестве меры влияния фактора на частоту события используется отношение шансов (Odds ratio, OR).

Логистическая регрессия отличается от обычной множественной регрессии тем, что зависимая переменная не является непрерывной, а может принимать только два значения.

Классический дискриминантный анализ является оптимальным методом анализа бинарного отклика в случае, когда выполнено основное условие его применения: данные получены из двух многомерных нормальных распределений с равными ковариационными матрицами, что на практике наблюдается достаточно редко. В остальных случаях применяется другой метод дискриминантного анализа – логистическая регрессия.

Основой алгоритма классификации является дискриминантная функция, результат вычисления которой для каждого наблюдения сравнивается с заданным порогом разделения массива.

В классическом дискриминантном анализе результат вычисления дискриминантной функции может находиться в диапазоне от -∞ до+∞.

В логистической регрессии осуществляется логит-преобразование дискриминантной функции, трансформирующее результат её вычисления для каждого наблюдения в вероятность, ограниченную диапазоном от 0 до 1.

Логистическая регрессия не предъявляет жестких требований к характеру распределения и объему выборки. Порогом классификации является уровень вероятности p=0,5, который соответствует отношению шансов OR=1.

2.2. Математическое обоснование

Логистическая регрессия – статистическая модель прогнозирования вероятности события по значениям предикторов, в которой зависимая переменная принимает лишь одно из двух значений: 0 –событие не произошло и 1 –событие произошло.

Уравнение логистической регрессии имеет вид:

$$P(Y) = \frac{1}{1 + e^{-z}}$$

где Y – бинарная зависимая переменная, принимающая только два значения (1 и 0),

P – условная вероятность того, что Y примет значение, равное единице, P {Y=1 | x}=f(z),

z – дискриминантная функция,

$z = \beta 0 + \beta_1 x_1 + \ldots + \beta_2 x_2 + \ldots + \beta_i x_i$ – векторы значений независимых переменных x_i и коэффициентов регрессии β_i. β_0 – константа.

Логистическая регрессия представляет собой такое преобразование линейной регрессии, при котором зависимая переменная ограничена диапазоном от 0 до 1 и предсказывает вероятность события в зависимости от значений предикторов. Подобного вида трансформация осуществляется с помощью логистической функции, уравнение и график которой представлены на рис. 6.

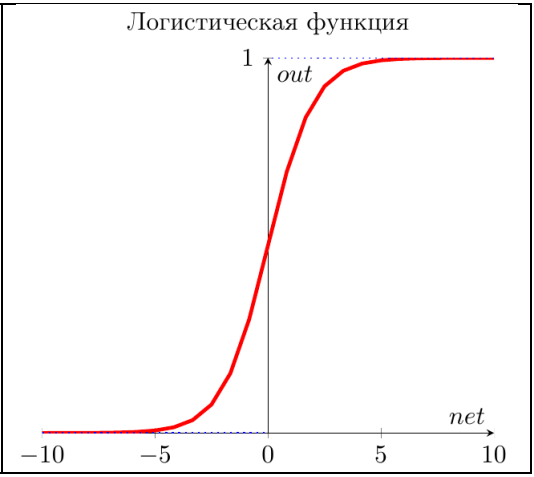
$f(z) = \frac{1}{1+e^{-z}}$	

Рис. 6. Уравнение и график логистической функции *f(z)* [оригинальный рисунок].

2.3. Взаимосвязь логистической функции и вероятности

Использование логистической функции для преобразования линейной регрессии в логистическую имеет под собой глубокое обоснование, поскольку логистическая функция является функцией, обратной *логит-функции*, представляющей собой натуральный логарифм отношения шансов:

$$\text{logit}\,(p) = \ln(OR)$$

Поскольку шансы и вероятность связаны между собой, появляется возможность вычислять вероятность принадлежности объекта к определенному классу:

$$\text{logit (p)} = \ln(OR) = \ln\left(\frac{p(X)}{1-p(X)}\right) = z, \qquad (1)$$

где z – векторы значений независимых переменянных и коэффициентов регрессии

$$z = \beta_0 + \beta_1 x_1 + + \beta_2 x_2 + \ldots + \beta_i x_i .$$

Если взять экспоненты от обеих частей уравнения (1), получим

$$\frac{p(X)}{1-p(X)} = e^z \text{, откуда}$$

$$p(X) = \frac{e^z}{(1-e^z)} = \frac{1}{(1+e^{-z})}$$

В окончательном виде уравнение логистической регрессии приобретает вид:

$$p(X) = \frac{1}{\left(1+e^{-(\beta 0+\beta 1 X 1+\cdots+\beta i X i)}\right)} \qquad (2)$$

2.4. Условия применения и выбор предикторов

1. Модель логистической регрессии может применяться в качестве объясняющей для оценки степени влияния предикторов на зависимую переменную, а также в качестве прогностической для классификации новых наблюдений. В последнем случае построенная модель нуждается в проверке её работы (валидации) на независимой выборке объектов. Если число наблюдений достаточно велико, исходная выборка может

быть разделена на обучающую и тестовую, отобранную случайным образом. Совпадение параметров моделей для основной и тестовой выборок является признаком устойчивости, и модель может быть успешно использована для классификации новых объектов.

2. В качестве предикторов могут включаться переменные численного, интервального и категориального типа. Число предикторов не должно превышать 1/10 количества наблюдений.

3. Отбор предикторов для включения в мультивариантную модель можно проводить по уровню статистической значимости <0,1 по результатам ROC-анализа.

4. Должна быть исследована взаимная корреляция переменных. В качестве предиктора при наличии сильной корреляции (R>0,7) следует использовать лишь одну из них. Выбор переменной должен осуществляться, исходя из логики рассуждений о характере возможной взаимосвязи, проверяемой гипотезы и здравого смысла. Так, например, сильно коррелированными являются толщина меланомы по Бреслоу, уровень инвазии по Кларку и Т-категория опухоли по TNM. В подобных случаях могут быть построены и оценены различные модели регрессии, включающие в число предикторов одну из перечисленных переменных.

Глава 3.
Логистическая регрессия в MedCalc

Путь к модулю логистической регрессии: Statistics/Regression/Logistic regression (рис. 7).

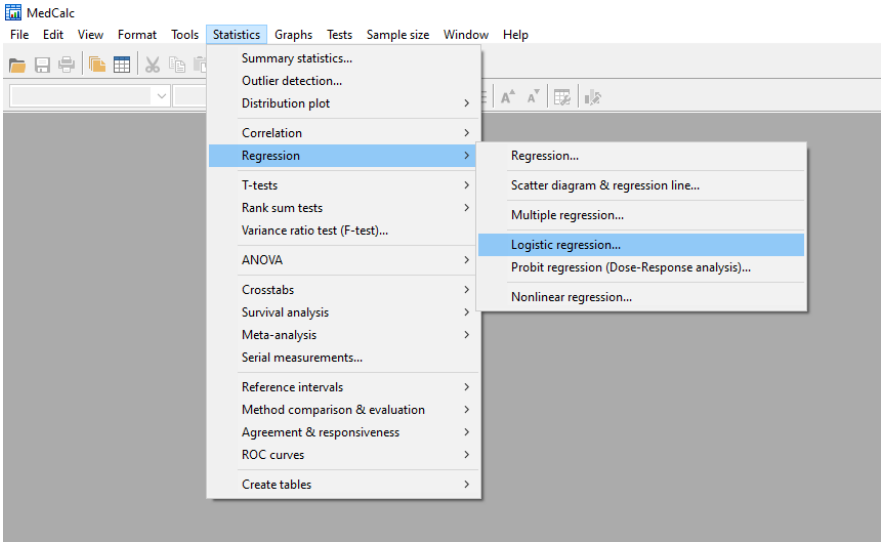


Рис. 7. Путь к модулю Logistic regression [оригинальный рисунок].

Для иллюстрации работы модуля рассмотрим пример построения модели логистической регрессии оценки эффективности изолированной химиоперфузии легкого (рис. 8).

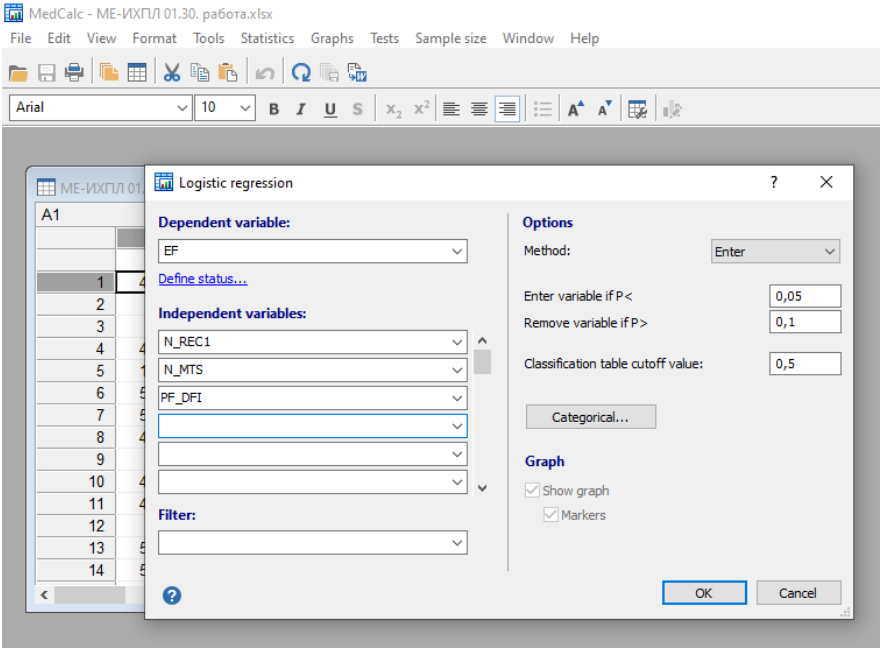


Рис. 8. Выбор переменных и способа включения в мультивариантную логистическую модель [оригинальный рисунок].

Dependent variable

Критерием эффективности лечения является зависимая переменная EF, принимающая два значения:

1 – эффект лечения есть,

0 – эффекта лечения нет.

Положительным эффектом в проводимом исследовании считали удлинение легочного безрецидивного интервала более чем в 1,5 раза у одних и тех же пациентов, перенесших и стандартную метастазэктомию и метастазэктомию с химиоперфузией.

Independent variables

В качестве предикторов вводятся отобранные в ходе ROC-анализа по уровню значимости p<0,1 независимые переменные:

N_REC1 – число рецидивных легочных метастазов,

N_MTS – число легочных метастазов при первичной операции,

PF_DFI – длительность легочного безрецидивного промежутка.

Filter

Фильтр предназначается для проведения анализа в конкретной подгруппе, принадлежность к которой обозначается отдельной группирующей переменной (в нашем примере не используется).

Options. Опции:

Method

Метод ввода – способ введения в модель независимых переменных

• Enter: ввод всех переменных в модель за один шаг без проверки.

• Forward (вперед) – последовательное введение значимых переменных.

• Backward (назад) – введение всех переменных с последовательным удалением незначимых.

• Stepwise (пошаговый) – последовательное введение значимых переменных с проверкой уровня значимости и возможным удалением переменных, которые стали незначимыми.

• Enter variable if P < – Ввести переменную, если связанный с ней уровень значимости меньше заданного Р-значения.

- Remove variable if P > – Удалить переменную, если связанный с ней уровень значимости больше данного P-значения.

Graf

Опция построения графика, отображающего кривую логистической регрессии, в MedCalc доступна только при наличии одной единственной независимой переменной.

3.1. Оценка качества логистической регрессионной модели в MedCalc

Программа генерирует детальный список характеристик модели. В таблице 3 представлены две модели, одна из которых включает все указанные переменные, а в другой исключена переменная N_MTS, получившая низкий уровень значимости в предыдущей модели 1.

Overall Model Fit. Общее соответствие модели.

Нулевая модель Null Model -2 Log Likelihood задается -2ln(L0), где L0 – вероятность получения результатов наблюдений, в случае, если независимые переменные не оказывают влияния на результат.

Полная модель Null Model -2 Log Likelihood задается -2 ln(L), где L – вероятность получения результатов наблюдений при включении всех независимых переменных в модель.

Разность этих двух значений дает статистику Chi-Squared, которая является мерой того, насколько сильно независимые переменные влияют на зависимую переменную.

Сравнение моделей логистической регрессии, построенных в MedCalc, представлено в таблице 3.

Significance level – уровень статистической значимости модели. Если P-значение для общей статистики соответствия модели <0,05, то хотя бы одна из независимых переменных влияет на результат.

Cox & Snell R2 и Nagelkerke R2 – это другие показатели хорошего соответствия, известные как псевдо R-квадраты. ПсевдоR-квадрат Кокса и Снелла имеет максимальное значение, которое не равно 1. Нагелькерке R2 корректирует показатель Кокса и Снелла таким образом, чтобы диапазон возможных значений расширялся до 1.

Коэффициенты логистической регрессии показывают изменение (увеличение при βi>0, уменьшение при βi<0) прогнозируемой логарифмированной вероятности наличия интересующей характеристики при изменении независимых переменных на одну единицу.

Если независимые переменные *Xa* и *Xb* являются дихотомическими переменными, то влияние этих переменных на зависимую переменную можно оценить, сравнивая их коэффициенты регрессии *βa* и *βb*.

Таблица 3

Сравнение моделей логистической регрессии,
построенных в MedCalc [оригинальная таблица]

Logistic regression Model 1

Dependent Y	EF
Method	Enter
Sample size	42
Positive cases [a]	32 (76,19%)
Negative cases [b]	10 (23,81%)

[a] EF = 1
[b] EF = 0

Overall Model Fit

Null model -2 Log Likelihood	46,105
Full model -2 Log Likelihood	28,669
Chi-squared	17,436
DF	3
Significance level	P = 0,0006
Cox & Snell R^2	0,3398
Nagelkerke R^2	0,5099

Coefficients and Standard Errors

Variable	Coefficient	Std. Error	Wald	P
N_REC1	-0,76288	0,31991	5,6866	0,0171
N_MTS	-0,19584	0,20119	0,9475	0,3304
PF_DFI	0,21588	0,098920	4,7629	0,0291
Constant	-0,32447	0,88599	0,1341	0,7142

Logistic regression Model 2

Dependent Y	EF
Method	Enter
Sample size	44
Positive cases [a]	32 (72,73%)
Negative cases [b]	12 (27,27%)

[a] EF = 1
[b] EF = 0

Overall Model Fit

Null model -2 Log Likelihood	51,564
Full model -2 Log Likelihood	30,723
Chi-squared	20,841
DF	2
Significance level	P < 0,0001
Cox & Snell R^2	0,3773
Nagelkerke R^2	0,5466

Coefficients and Standard Errors

Variable	Coefficient	Std. Error	Wald	P
N_REC1	-0,80302	0,30901	6,7530	0,0094
PF_DFI	0,22917	0,093665	5,9864	0,0144
Constant	-1,00185	0,75553	1,7583	0,1848

Odds Ratios and 95% Confidence Intervals

Variable	Odds ratio	95% CI
N_REC1	0,4663	0,2491 to 0,8730
N_MTS	0,8221	0,5542 to 1,2196
PF_DFI	1,2410	1,0222 to 1,5065

Hosmer & Lemeshow test

Chi-squared	8,6092
DF	9
Significance level	P = 0,4741

Classification table (cut-off value p=0,5)

Actual group	Predicted group		Percent correct
	0	1	
Y = 0	6	4	60,00%
Y = 1	4	28	87,50%
Percent of cases correctly classified			80,95%

ROC curve analysis

Area under the ROC curve (AUC)	**0,878**
Standard Error	0,0544
95% Confidence interval	0,740 to 0,959

Odds Ratios and 95% Confidence Intervals

Variable	Odds ratio	95% CI
N_REC1	0,4480	0,2445 to 0,8209
PF_DFI	1,2576	1,0466 to 1,5110

Hosmer & Lemeshow test

Chi-squared	3,3693
DF	9
Significance level	P = 0,9478

Classification table (cut-off value p=0,5)

Actual group	Predicted group		Percent correct
	0	1	
Y = 0	8	4	66,67%
Y = 1	4	28	87,50%
Percent of cases correctly classified			81,82%

ROC curve analysis

Area under the ROC curve (AUC)	**0,893**
Standard Error	0,0478
95% Confidence interval	0,763 to 0,966

Wald

Статистика Вальда – это коэффициент регрессии, деленный на квадрат стандартной ошибки: β/SE^2. P- уровень значимости критерия Вальда: переменная значима, если этот уровень меньше заданной величины (обычно 0,05).

Уровень статистической значимости модели 2 выше, так же, как и AUC (площадь под ROC кривой). Модель 2 имеет более высокую точность и правильно классифицирует 81,8% наблюдений. Тест Хосмера-Лемешоу *(Hosmer & Lemeshow Test)* указывает на достаточную калибровку обеих моделей. Исходя из перечисленного, предпочтение следует отдать модели 2.

Тест Хосмера-Лемешоу – это статистический тест на качество соответствия модели логистической регрессии. Данные делятся примерно на десять групп, определенных в порядке возрастания предполагаемого риска. Подсчитывается наблюдаемое и ожидаемое количество случаев в каждой группе и рассчитывается статистика χ-квадрат. Большое значение χ-квадрат (при малом $p < 0,05$) указывает на плохое соответствие, а малые значения χ-квадрат (при большем p-значении ближе к 1) указывают на хорошее соответствие модели логистической регрессии. В таблице Contingency Table for Hosmer and Lemeshow Test показаны детали теста с наблюдаемым и ожидаемым количеством случаев в каждой группе.

Odds Ratios and 95% Confidence Intervals. Отношение шансов и 95% доверительный интервал.

$$ln(OR) = \beta_0 + \beta_1 x_1 + + \beta_2 x_2 + \ldots + \beta_i x_i$$

Взяв экспоненту обеих сторон уравнения регрессии, как указано выше, уравнение можно переписать следующим образом:

$$OR = e^{(\beta 0 + \beta 1 x 1 + + \beta 2 x 2 + \cdots + \beta i x i)}$$

$$OR = e^{\beta 0} * e^{\beta 1 x 1} * e^{\beta 2 x 2} * \ldots . * e^{\beta i x i}$$

Очевидно, что если переменная x_i увеличивается на 1 единицу, а все остальные факторы остаются неизменными, то шансы *(OR)* увеличиваются в $e^{\beta i}$ раз.

Этот коэффициент $e^{\beta i}$ является «скорректированным» отношением шансов *(OR)* для независимой переменной x_i и дает относительную величину, на которую увеличивается ($OR>1$) или уменьшается $(OR<1)$ вероятность исхода при увеличении значения независимой переменной на 1 единицу.

В нашем примере (модель 2) переменная N_REC1 кодирует число рецидивных узлов в легких, развившихся после метастазэктомии с химиоперфузией. Отношение шансов для этой переменной равно 0,448. Это означает, что в данной модели шансы на положительный исход с увеличением на единицу числа рецидивных узлов снижаются на 0,448.

3.2. Интерпретация уравнения логистической регрессии

Уравнение логистической регрессии в модели 2 имеет вид:

logit(p)= −1,00185 - 0,80302*(N_REC) + 0,922817*(PF_DFI)

Таким образом, для пациента с одним метастатическим узлом (N_REC1=1), развившимся через 12 месяцев после химиоперфузии (PF_DFI=12) logit(p) равен 0,946.

Logit(p) может быть обратно преобразован в p по следующей формуле:

$$p = \frac{1}{1 + e^{-logit(p)}}$$

Для logit(p)=0,946 вероятность p положительного исхода составляет 0,280, т.е. эффект химиоперфузии будет отрицательным.

Classification Table

Таблица классификации – это еще один метод оценки прогностической точности модели логистической регрессии. В этой таблице перекрестно классифицируются наблюдаемые значения зависимого исхода и прогнозируемые значения (при заданном пользователем значении отсечения, например, p=0,50). В нашем примере модель 1 правильно предсказывает 80,95% случаев, модель 2 – 81,82%.

ROC curve analysis

Анализ ROC-кривой. Другой метод оценки модели логистической регрессии использует анализ ROC-кривой. Способность модели различать положительные и отрицательные случаи измеряется площадью под ROC-кривой (AUC). AUC, иногда называемая C-статистикой (или индексом конкордации), представляет собой значение, которое варьирует от 0,5 (дискриминационная способность отсутствует) до 1,0 (идеальная дискриминационная способность).

Чтобы выполнить полный анализ ROC-кривой на предсказанных вероятностях, можно сохранить предсказанные вероятности в качестве новой переменной и затем использовать её в анализе ROC-кривой. Зависимая переменная, использованная в логистической регрессии, затем выступает в качестве классификационной переменной в диалоговом окне анализа ROC-кривой.

В общем случае может быть рекомендовано введение в модель всех переменных с последовательным исключением не получивших достаточного уровня статистической значимости (метод пошагового исключения), однако иногда дополнительное введение предиктора с низкой статистической значимостью может заметно улучшать качество модели.

Контрольные вопросы

1. Что такое ROC-кривая?
2. Что такое регрессия?
3. В чем различия между каноническим дискриминантным анализом и логистической регрессией?
4. Что такое чувствительность и специфичность алгоритма классификации?
5. Что такое порог отсечения классификатора?
6. Что такое индекс Юдена?
7. Что такое логистическая функция?
8. Что такое логит-функция?
9. В чем отличие объясняющей и предсказательной моделей?
10. Что такое уровень статистической значимости модели?

Тестовые задания

Инструкция: выберите один или несколько правильных ответов.

1. ROC-кривая – это:

Поле для выбора ответа	Варианты ответов	Поле для отметки правильного ответа (+)
а	график, отражающий зависимость чувствительности классификатора от его специфичности	
б	график, отражающий зависимость специфичности классификатора от его чувствительности	
в	график, отражающий зависимость чувствительности классификатора от его параметра (1 – специфичность)	+
г	график, отражающий зависимость числа истинно положительных ответов классификатора от числа ложноотрицательных ответов	+

2. Из нескольких классификаторов лучшим считают классификатор:

Поле для выбора ответа	Варианты ответов	Поле для отметки правильного ответа (+)
а	имеющий наибольшую чувствительность	
б	имеющий наибольшую специфичность	
в	имеющий наиболее высокий уровень статистической значимости	
г	обеспечивающий наибольшую площадь под ROC-кривой (AUC) при высоком уровне значимости	+

3. Ограничением к интраоперационной рентгенографии образцов является следующий тип плотности тканей МЖ:

Поле для выбора ответа	Варианты ответов	Поле для отметки правильного ответа (+)
а	A	
б	B	
в	C	
г	D	+

3. В моделях логистической регрессии могут быть использованы следующие форматы данных:

Поле для выбора ответа	Варианты ответов	Поле для отметки правильного ответа (+)
а	текстовый	
б	категориальный	+
в	числовой	+
г	логический	
д	интервальный	+

4. Требования к независимым переменным в моделях логистической регрессии:

Поле для выбора ответа	Варианты ответов	Поле для отметки правильного ответа (+)
а	нормальный характер распределения	
б	равенство дисперсий	
в	взаимокорреляция $r>0,7$	
г	взаимокорреляция $r<0,7$	+
д	взаимокорреляция $r<0,5$	
е	взаимокорреляция $r<0,3$	

5. Требования к числу независимых переменных в моделях логистической регрессии:

Поле для выбора ответа	Варианты ответов	Поле для отметки правильного ответа (+)
а	отсутствуют	
б	менее 1/15 количества наблюдений	
в	менее 1/10 количества наблюдений	+
г	менее 1/5 количества наблюдений	

6. Отбор предикторов для включения в модель логистической регрессии осуществляется по уровню статистической значимости в ROC-анализе:

Поле для выбора ответа	Варианты ответов	Поле для отметки правильного ответа (+)
а	P <0,15	
б	P <0,1	+
в	P <0,05	
г	P <0,01	

7. Статистика Вальда – это:

Поле для выбора ответа	Варианты ответов	Поле для отметки правильного ответа (+)
а	квадрат коэффициента регрессии:	
б	β^2коэффициент регрессии, деленный на величину стандартной ошибки: β/SE	
в	коэффициент регрессии, деленный на квадрат стандартной ошибки: β/SE^2	+
г	коэффициент регрессии, деленный на число наблюдений: β/n	

8. В модель логистической регрессии статистически значимые переменные вводятся:

Поле для выбора ответа	Варианты ответов	Поле для отметки правильного ответа (+)
а	произвольно	
б	методом пошагового исключения	+
в	методом пошагового добавления	+
г	методом случайной подстановки	

Список литературы

1. Григорьев С.Г., Лобзин Ю.В., Скрипниченко Н.В. Роль и место логистической регрессии и ROC-анализа в решении медицинских диагностических задач // Журнал инфектологии. – 2016. – Т.8, № 4. – С. 36-43.

2. Каримов Р.Н. Статистика для врачей, биологов и не только. – в 2 ч. / Р.Н. Каримов, Ю.Г. Шварц. – Саратов: Сарат. гос. мед. ун-т, 2007. – Ч.1. – 200 с.; 2010. – Ч.2. – 204 с.

3. Hosmer D.W. Applied Logistic Regression. – 2-nd ed. / D.W. Hosmer, S. Lemeshow. – N.-Y.: Wiley, 2000. – 375 c.

4. https://www.medcalc.org/MedCalc statistical software manual. Version 17 of the software.

ISBN 978-5-6048249-5-5

9 785604 824955

Отпечатано в ООО «АРТЕК»,
СПб, 6-я линия В.О., д. 3/10
E-mail: artek-1@mail.ru, т. +7(911) 239-25-32
Подписано в печать 01.06.23
Формат 60x90/16. Тираж 50 экз.