**Задание 7\_3. Бинарная регрессия R**

Логистическая регрессия — это разновидность [множественной регрессии](https://wiki.loginom.ru/articles/multiple-linear-regression.html), общее назначение которой состоит в анализе связи между несколькими [независимыми переменными](https://wiki.loginom.ru/articles/input-variable.html) (называемыми также регрессорами или предикторами) и [зависимой переменной](https://wiki.loginom.ru/articles/output-variable.html). Бинарная логистическая регрессия применяется в случае, когда зависимая переменная является [бинарной](https://wiki.loginom.ru/articles/binary-variable.html) (т.е. может принимать только два значения). С помощью логистической регрессии можно оценивать вероятность того, что [событие](https://wiki.loginom.ru/articles/random-event.html) наступит для конкретного испытуемого (больной/здоровый, возврат кредита/дефолт и т.д.).

Модель логистической регрессии имеет следующий вид:



где *xi*– предикторы, *bi* – коэффициенты регрессии.

Пример.

Вводим данные по результатам приема абитуриента в университет (принят или нет). Три независимые переменные: gre – оценка абитуриента на экзамене; gpa – средний балл; rank – престиж школы, в которой обучался абитуриент (1 – самая престижная, 4 – наименее престижная). Результат приема (admit) – бинарная переменная.

mydata <- read.csv("http://www.ats.ucla.edu/stat/data/binary.csv")

Функция glm позволяет строить логит модели бинарной регрессии. Сначала преобразуем зависимую переменную в категориальную.

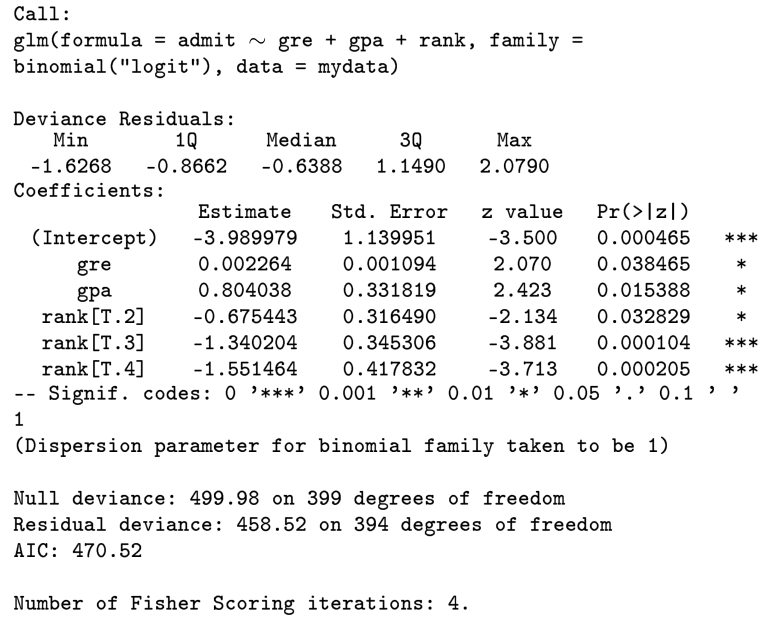
mydata$rank <- factor(mydata$rank)

Н0: Между переменными не существует логистической зависимости.

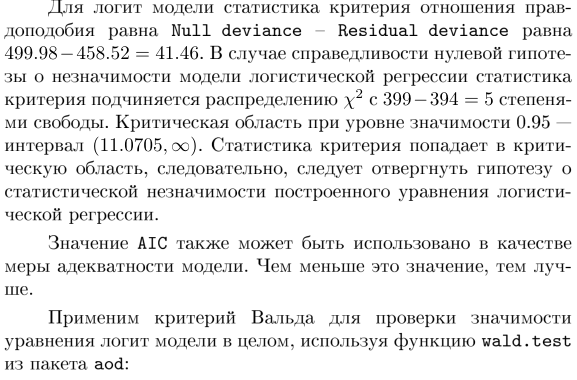
mylogit <- glm(admit ∼ gre + gpa + rank, data = mydata, family= binomial("logit"))

Для вывода результатов используем следующую функцию:

summary(mylogit); AIC(mylogit) # Получим результаты:



Все переменные в модели значимы, т.к р-value<0.05.



install.packages("aod")

library(aod)

wald.test(b = coef(mylogit), Sigma = vcov(mylogit), Terms =1:6)

Wald test:

Chi-squared test:

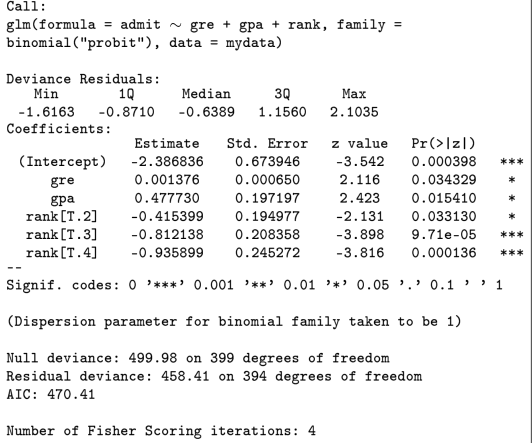
X2 = 73.4, df = 6, P(> X2) = 8.2e-14

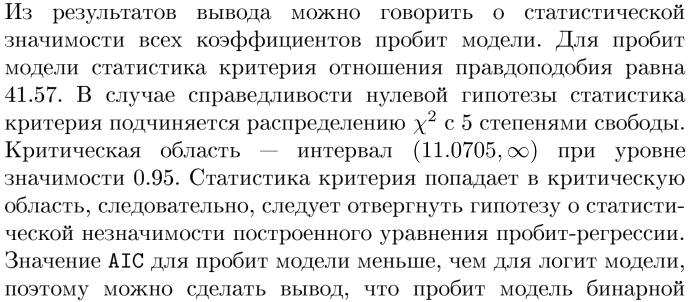
Так как, р-value=8.2e-14<0.05, то уравнение логистической регрессии значимо в целом.

Аналогичным образом можно построить **модель пробит-регрессии**.

myprobit <- glm(admit ~gre + gpa + rank, data = mydata, family= binomial("probit"))

summary(myprobit); AIC(myprobit)









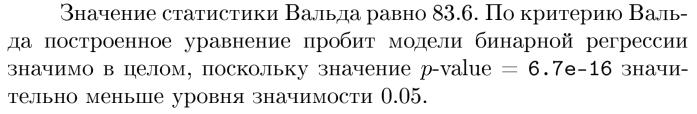
library(aod)

wald.test(b = coef(myprobit), Sigma = vcov(myprobit), Terms =1:6)

Wald test:

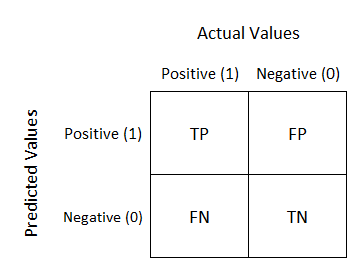
Chi-squared test:

X2 = 83.6, df = 6, P(> X2) = 6.7e-16



Для проверки адекватности бинарной модели можно рассчитать среднеквадратическую ошибку и построить матрицу неточности (функция confusionMatrix() из пакета caret).

Матрица неточностей или ошибок - это таблица с 4 различными комбинациями прогнозируемых и фактических значений.



где TP — истино-положительное решение, TN — истино-отрицательное решение, FP — ложно-положительное решение (Ошибка типа 1), FN— ложно-отрицательное решение (Ошибка Типа 2):

Пример матрицы неточностей

|  |  |  |
| --- | --- | --- |
| Наблюдаемые значения | Прогноз | |
| 0 | 1 |
| 0 | 25 | 0 |
| 1 | 2 | 113 |

Два объекта неправильно предсказаны.

Оценка качества модели (accuracy) рассчитывалась как доля правильно классифицированных объектов (количество правильно классифицированных объектов к общему числу объектов):

.

Чувствительность модели (Sensitivity) равна проценту верно предсказанных позитивных исходов. Ее рассчитывают по матрице неточности: (113/(113+2))\*100%=98.26%. Специфичность модели (Specificity) показывает процент верно предсказанных негативных исходов: (25/(25+0))\*100%= 100%. Построенная модель на 100% предсказывает отсутствие положительной динамики при лечении больных туберкулезом в стационаре и на 98.26% - наличие положительной динамики.

**Задание**.

Откройте таблицу данных datlg (медицинские данные по кардиологии: y – зависимая переменная – стоит делать операцию пациенту или нет в зависимости от имеющихся предикторов).

Нужно построить логит и пробит модели и сравнить полученные результаты.

## Таблица критические точки распределения хи-квадрат, где α — уровень значимости α (вероятность), k — количество степеней свободы.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **k** | **Уровень значимости α** | | | | | | | | | | | | |
| **0,005** | **0,01** | **0,025** | **0,05** | **0,1** | **0,25** | **0,5** | **0,75** | **0,9** | **0,95** | **0,975** | **0,99** | **0,995** |
| **1** | *7,87944* | *6,63490* | *5,02389* | *3,84146* | *2,70554* | *1,32330* | *0,45494* | *0,10153* | *0,01579* | *0,00393* | *0,00098* | *0,00016* | *0,00004* |
| **2** | *10,59663* | *9,21034* | *7,37776* | *5,99146* | *4,60517* | *2,77259* | *1,38629* | *0,57536* | *0,21072* | *0,10259* | *0,05064* | *0,02010* | *0,01003* |
| **3** | *12,83816* | *11,34487* | *9,34840* | *7,81473* | *6,25139* | *4,10834* | *2,36597* | *1,21253* | *0,58437* | *0,35185* | *0,21580* | *0,11483* | *0,07172* |
| **4** | *14,86026* | *13,27670* | *11,14329* | *9,48773* | *7,77944* | *5,38527* | *3,35669* | *1,92256* | *1,06362* | *0,71072* | *0,48442* | *0,29711* | *0,20699* |
| **5** | *16,74960* | *15,08627* | *12,83250* | *11,07050* | *9,23636* | *6,62568* | *4,35146* | *2,67460* | *1,61031* | *1,14548* | *0,83121* | *0,55430* | *0,41174* |
| **6** | *18,54758* | *16,81189* | *14,44938* | *12,59159* | *10,64464* | *7,84080* | *5,34812* | *3,45460* | *2,20413* | *1,63538* | *1,23734* | *0,87209* | *0,67573* |
| **7** | *20,27774* | *18,47531* | *16,01276* | *14,06714* | *12,01704* | *9,03715* | *6,34581* | *4,25485* | *2,83311* | *2,16735* | *1,68987* | *1,23904* | *0,98926* |
| **8** | *21,95495* | *20,09024* | *17,53455* | *15,50731* | *13,36157* | *10,21885* | *7,34412* | *5,07064* | *3,48954* | *2,73264* | *2,17973* | *1,64650* | *1,34441* |
| **9** | *23,58935* | *21,66599* | *19,02277* | *16,91898* | *14,68366* | *11,38875* | *8,34283* | *5,89883* | *4,16816* | *3,32511* | *2,70039* | *2,08790* | *1,73493* |
| **10** | *25,18818* | *23,20925* | *20,48318* | *18,30704* | *15,98718* | *12,54886* | *9,34182* | *6,73720* | *4,86518* | *3,94030* | *3,24697* | *2,55821* | *2,15586* |
| **11** | *26,75685* | *24,72497* | *21,92005* | *19,67514* | *17,27501* | *13,70069* | *10,34100* | *7,58414* | *5,57778* | *4,57481* | *3,81575* | *3,05348* | *2,60322* |
| **12** | *28,29952* | *26,21697* | *23,33666* | *21,02607* | *18,54935* | *14,84540* | *11,34032* | *8,43842* | *6,30380* | *5,22603* | *4,40379* | *3,57057* | *3,07382* |
| **13** | *29,81947* | *27,68825* | *24,73560* | *22,36203* | *19,81193* | *15,98391* | *12,33976* | *9,29907* | *7,04150* | *5,89186* | *5,00875* | *4,10692* | *3,56503* |
| **14** | *31,31935* | *29,14124* | *26,11895* | *23,68479* | *21,06414* | *17,11693* | *13,33927* | *10,16531* | *7,78953* | *6,57063* | *5,62873* | *4,66043* | *4,07467* |
| **15** | *32,80132* | *30,57791* | *27,48839* | *24,99579* | *22,30713* | *18,24509* | *14,33886* | *11,03654* | *8,54676* | *7,26094* | *6,26214* | *5,22935* | *4,60092* |
| **16** | *34,26719* | *31,99993* | *28,84535* | *26,29623* | *23,54183* | *19,36886* | *15,33850* | *11,91222* | *9,31224* | *7,96165* | *6,90766* | *5,81221* | *5,14221* |
| **17** | *35,71847* | *33,40866* | *30,19101* | *27,58711* | *24,76904* | *20,48868* | *16,33818* | *12,79193* | *10,08519* | *8,67176* | *7,56419* | *6,40776* | *5,69722* |
| **18** | *37,15645* | *34,80531* | *31,52638* | *28,86930* | *25,98942* | *21,60489* | *17,33790* | *13,67529* | *10,86494* | *9,39046* | *8,23075* | *7,01491* | *6,26480* |
| **19** | *38,58226* | *36,19087* | *32,85233* | *30,14353* | *27,20357* | *22,71781* | *18,33765* | *14,56200* | *11,65091* | *10,11701* | *8,90652* | *7,63273* | *6,84397* |
| **20** | *39,99685* | *37,56623* | *34,16961* | *31,41043* | *28,41198* | *23,82769* | *19,33743* | *15,45177* | *12,44261* | *10,85081* | *9,59078* | *8,26040* | *7,43384* |
| **21** | *41,40106* | *38,93217* | *35,47888* | *32,67057* | *29,61509* | *24,93478* | *20,33723* | *16,34438* | *13,23960* | *11,59131* | *10,28290* | *8,89720* | *8,03365* |
| **22** | *42,79565* | *40,28936* | *36,78071* | *33,92444* | *30,81328* | *26,03927* | *21,33705* | *17,23962* | *14,04149* | *12,33801* | *10,98232* | *9,54249* | *8,64272* |
| **23** | *44,18128* | *41,63840* | *38,07563* | *35,17246* | *32,00690* | *27,14134* | *22,33688* | *18,13730* | *14,84796* | *13,09051* | *11,68855* | *10,19572* | *9,26042* |
| **24** | *45,55851* | *42,97982* | *39,36408* | *36,41503* | *33,19624* | *28,24115* | *23,33673* | *19,03725* | *15,65868* | *13,84843* | *12,40115* | *10,85636* | *9,88623* |
| **25** | *46,92789* | *44,31410* | *40,64647* | *37,65248* | *34,38159* | *29,33885* | *24,33659* | *19,93934* | *16,47341* | *14,61141* | *13,11972* | *11,52398* | *10,51965* |
| **26** | *48,28988* | *45,64168* | *41,92317* | *38,88514* | *35,56317* | *30,43457* | *25,33646* | *20,84343* | *17,29189* | *15,37916* | *13,84391* | *12,19815* | *11,16024* |
| **27** | *49,64492* | *46,96294* | *43,19451* | *40,11327* | *36,74122* | *31,52841* | *26,33634* | *21,74941* | *18,11390* | *16,15140* | *14,57338* | *12,87850* | *11,80759* |
| **28** | *50,99338* | *48,27824* | *44,46079* | *41,33714* | *37,91592* | *32,62049* | *27,33623* | *22,65716* | *18,93924* | *16,92788* | *15,30786* | *13,56471* | *12,46134* |
| **29** | *52,33562* | *49,58788* | *45,72229* | *42,55697* | *39,08747* | *33,71091* | *28,33613* | *23,56659* | *19,76774* | *17,70837* | *16,04707* | *14,25645* | *13,12115* |
| **30** | *53,67196* | *50,89218* | *46,97924* | *43,77297* | *40,25602* | *34,79974* | *29,33603* | *24,47761* | *20,59923* | *18,49266* | *16,79077* | *14,95346* | *13,78672* |

Источники:

1. <https://hranalytic.ru/kak-ponyat-matrica-nesootvetstvij-confusion-matrix/>
2. http://distrland.blogspot.com/2018/02/r.html