SAINT PETERSBURG STATE UNIVERSITY

Faculty of Applied Mathematics and Control Processes

Mathematical Game Theory and Statistical Decisions Department

Applied Statistics in R

Laboratory work № 6

Professor: Parilina Elena M.

Student: Orlov Ivan M., 19.M09-пу

Saint Petersburg

2020

Logit for y ~ x1 + x2 + x3 + x4 + x5

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.4571	-0.463	0.1563	0.4326	1.7521

Coefficients (significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1)

	Estimate	Std. Error	z value	Pr(>|z|)	Significance code
(Intercept)	-10.94228	3.8706	-2.827	0.0047	**
x1	0.45547	0.18012	2.529	0.0114	*
x2	0.80851	0.4359	1.855	0.0636	.
x3	-0.35588	0.41778	-0.852	0.3943	
x4	0.1262	0.42222	0.299	0.765	
x5	-0.03185	0.01985	-1.605	0.1086	

Null deviance: 75.934 on 57 degrees of freedom
Residual deviance: 39.647 on 52 degrees of freedom
AIC: 51.647

Number of Fisher Scoring iterations: 6

Wald test:
X2 = 13.3, df = 6, P(> X2) = 0.038

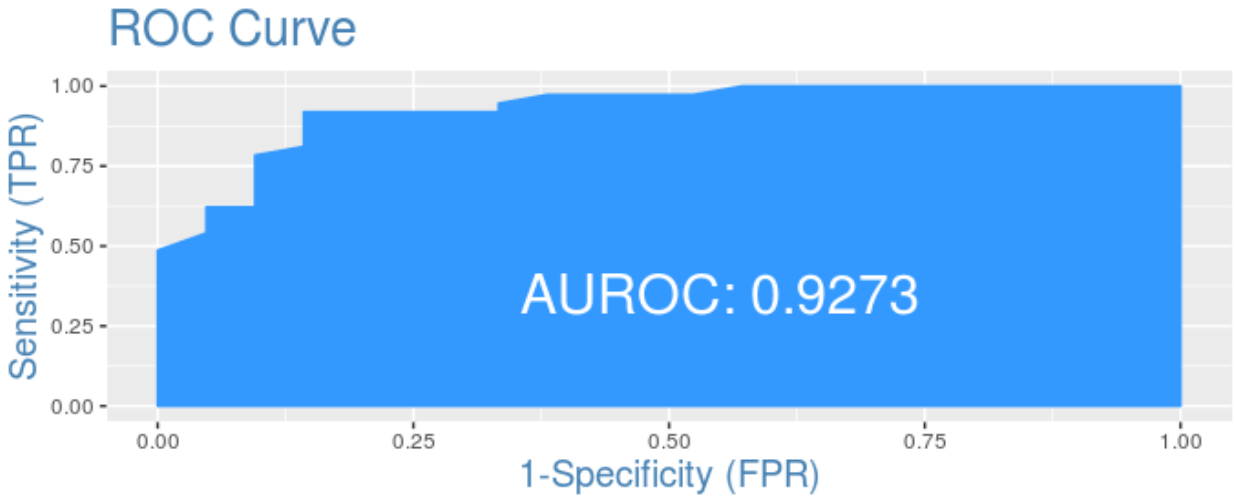
optimalCutoff = 0.4992771

Confusion matrix for default cutoff

	0	1
0	18	3
1	3	34

Confusion matrix for optimal cutoff

	0	1
0	18	3
1	3	34



Probit for y ~ x1 + x2 + x3 + x4 + x5

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.425	-0.4708	0.1202	0.4691	1.7234

Coefficients (significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1)

	Estimate	Std. Error	z value	Pr(>|z|)	Significance code
(Intercept)	-6.23863	2.05647	-3.034	0.00242	**
x1	0.26007	0.09729	2.673	0.00751	**
x2	0.46814	0.23433	1.998	0.04574	*
x3	-0.21734	0.23038	-0.943	0.34547	
x4	0.06027	0.23947	0.252	0.80128	
x5	-0.01792	0.01097	-1.634	0.10216	

Null deviance: 75.934 on 57 degrees of freedom
Residual deviance: 39.694 on 52 degrees of freedom
AIC: 51.694

Number of Fisher Scoring iterations: 7

Wald test:
X2 = 16.8, df = 6, P(> X2) = 0.01

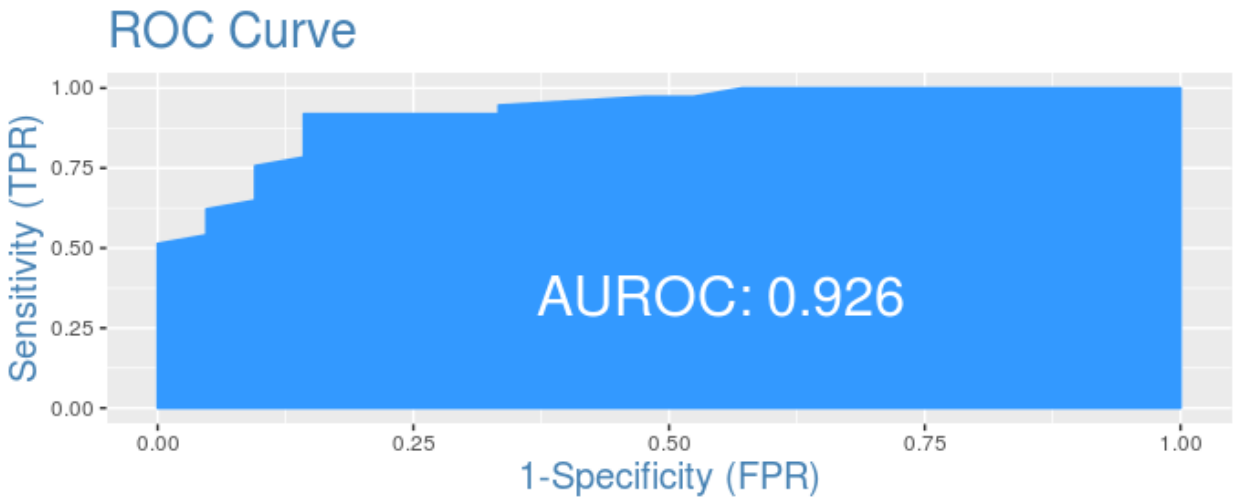
optimalCutoff = 0.5029279

Confusion matrix for default cutoff

	0	1
0	17	3
1	4	34

Confusion matrix for optimal cutoff

	0	1
0	18	3
1	3	34



For Doctor dataset logit regression model provides a little better result:

Deviance residuals are almost the same and close to 0.

Each of logit coefficients has bigger or equal level of significance than their probit vis-à-vis.

Both regressions are not significant on 0.05 level, but significant on 0.01.

Both optimal cutoffs are very close to default 0.5 yet when logit confusion matrix does not change with switch from default to optimal, probit reduces amount of false positives by 1.

AUROCs are basically the same with logit one being bigger by 0.0013.

Logit for Res ~ Sleep + Study

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.2246	-0.1425	0.0042	0.2007	2.1524

Coefficients (significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1)

Estimate	Std.	Error	z	value	Pr(>|z|)
(Intercept)	-16.1603	4.0819	-3.959	7.53E-05	***
Sleep	1.7454	0.4308	4.051	5.09E-05	***
Study	1.4865	0.3941	3.772	0.000162	***

Null deviance: 137.628 on 99 degrees of freedom
Residual deviance: 43.402 on 97 degrees of freedom
AIC: 49.402

Number of Fisher Scoring iterations: 8

Wald test:
X2 = 16.5, df = 3, P(> X2) = 0.00089

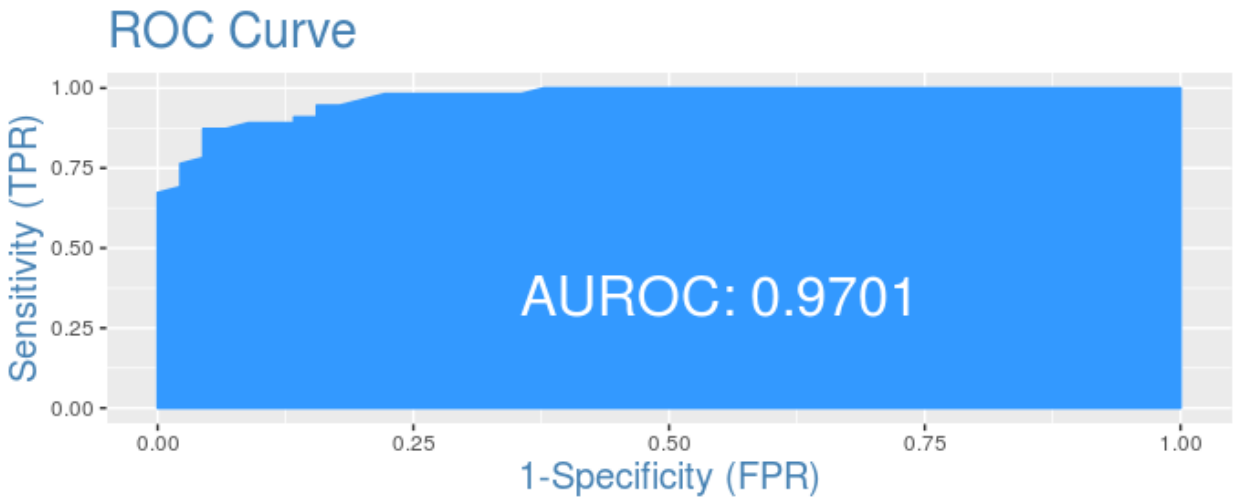
optimalCutoff = 0.6399997

Confusion matrix for default cutoff

	0	**1**
0	39	5
1	6	50

Confusion matrix for optimal cutoff

	0	**1**
0	43	7
1	2	48



Probit for Res ~ Sleep + Study

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.20524	-0.08993	0	0.15986	2.11765

Coefficients (significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1)

Estimate	Std.	Error	z	value	Pr(>|z|)
(Intercept)	-9.3426	2.1896	-4.267	1.98E-05	***
Sleep	1.0057	0.2291	4.39	1.13E-05	***
Study	0.8643	0.2155	4.012	6.03E-05	***

Null deviance: 137.628 on 99 degrees of freedom
Residual deviance: 43.074 on 97 degrees of freedom
AIC: 49.074

Number of Fisher Scoring iterations: 9

Wald test:
X2 = 19.4, df = 3, P(> X2) = 0.00022

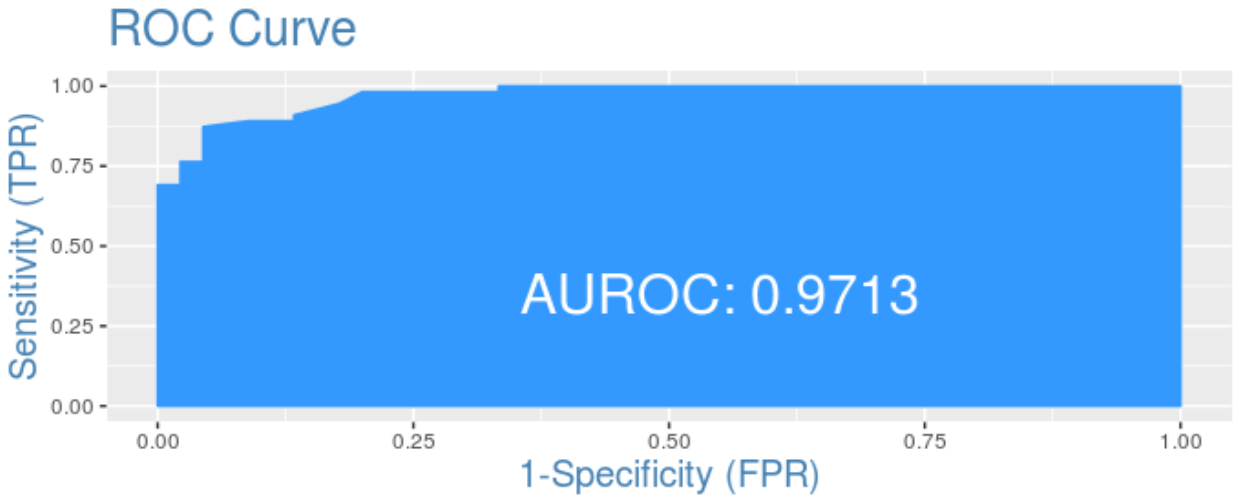
optimalCutoff = 0.5798182

Confusion matrix for default cutoff

	0	1
0	39	5
1	6	50

Confusion matrix for optimal cutoff

	0	1
0	43	7
1	2	48



For exam dataset probit regression model provides a little better result (based on AUROC):

Deviance residuals are almost the same and close to 0.

All coefficients of both models have significance level between 0 and 0.001levels.

Both regressions are not significant even on 0.0001 level.

Both logit and probit confusion matrices change witch from default to optimal cutoffs reducing amount of false positives by 4 but increasing false negatives by 2.

AUROCs are basically the same with probit being bigger by 0.0012.

```
library(readxl)
library(InformationValue)
library(aod)

doctor <- read_excel("Datasets/Doctor.xlsx")

exam <- read_excel("Datasets/binary regression.xls", col_names = FALSE)

names(exam) = c("Sleep", "Study", "Res")

research_bin = function(formula, data, actual_var, family_name){
  print(paste(family_name, "for", actual_var))
  mybin = glm(formula, data, family = binomial(family_name))
  print(summary(mybin))

  print(wald.test(b = coef(mybin), Sigma = vcov(mybin), Terms = 1:ncol(data)))

  predicted = plogis(predict(mybin, data))
  actual = data[[actual_var]]
  optCutOff = optimalCutoff(actual, predicted)[1]
  print(optCutOff)

  print(confusionMatrix(actual, predicted))
  print(confusionMatrix(actual, predicted, threshold = optCutOff))

  plotROC(actual, predicted)
}

datas = list(doctor, exam)

formulas = c(y ~ x1 + x2 + x3 + x4 + x5,
             Res ~ Sleep + Study)

actual_vars = c('y', 'Res')

family_names = list('logit', 'probit')

for (i in seq(1,2))
  for (f in family_names)
    research_bin(formulas[[i]], datas[[i]], actual_vars[[i]], f)
```