1 - Topic

Tools for the diagnostic and the assessment of logistic regression.

This tutorial describes the implementation of tools for the diagnostic and the assessment of a logistic regression. These tools are available in Tanagra version 1.4.33 (and later).

We deal with a credit scoring problem. We try to determine by using logistic regression the factors underlying the agreement or refusal of a credit to customers. We perform the following steps:

- Estimating the parameters of the classifier;
- Retrieving the covariance matrix of coefficients;
- Assessment using the Hosmer and Lemeshow goodness of fit test;
- Assessment using the reliability diagram;
- Assessment using the ROC curve;
- Analysis of residuals, detection of outliers and influential points.

On the one hand, we use **Tanagra 1.4.33**. Then, on the other hand, we perform the same analysis using the **R 2.9.2 software** [glm(.) procedure].

2 - Dataset

Our data file « LOGISTIC_REGRESSION_DIAGNOSTICS.XLS¹ » contains n = 100 observations. The binary target attribute is « ACCEPTATION.CREDIT » (« yes » or « no »). The predictive variables are

Name	Description	Туре
Age	Age of the customer	Continuous
Income.Per.Dependent	Income per dependent in the	Continuous
	household	
Derogatory.Report	At least one problem with the	Binary
	bank was reported	

3 - Analysis in Tanagra

Importing the data file

To import the dataset, we open the file into Excel spreadsheet. Then, by the way of Tanagra.xla² addin, we click on the TANAGRA / EXECUTE TANAGRA menu. Tanagra is automatically launched, the data file is imported.

¹ <u>http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/logistic_regression_diagnostics.zip</u>

² <u>http://data-mining-tutorials.blogspot.com/2008/10/excel-file-handling-using-add-in.html;</u> a similar tool is available for Open office Calc: <u>http://data-mining-tutorials.blogspot.com/2008/10/000calc-file-handling-using-add-in.html</u>

■ Fichier Edition Affichage Insertion Format Outils Données Fenêtre 2 Sipina Tanagra Adobe PDF □ <th>- 0 X</th>	- 0 X
A B C D E F G H About A1 • A Age A B C D E F G H About A1 • A Age C D E F G H About A1 • A Age Decoder Decoder Perspective About About 1 Age income.per.dependent Decogatory.reports Acceptation.Credit Acceptation Acceptatio	_ 8 ×
Bitstream Vera Sans • 8 • 6 I S E E E E About A1 • 6 I S E E E E About A1 • 6 A Age B C D E F G H 1 Age Income.per.dependent Derogatory reports Acceptation Credit Income.per.dependent Derogatory reports Income.per.dependent Derogatory reports Income.per.dependent Derogatory reports Income.per.dependent Derogatory reports Income.per.dependent Derogatory reports	
A1 • fx Age A B C D E F G H 1 Age Income.per.dependent Derogatory reports Acceptation Credit	
A B C D E F G H 1 Age Income.per.dependent Derogatory reports Acceptation.Credit	
1 Age Income per dependent Derogatory reports Acceptation Credit 2 38 4.52 0 yes 3 33 2.42 0 yes 4 34 4.5 0 yes 5 31 2.54 0 yes 6 32 9.79 0 yes 7 23 2.5 0 yes 9 29 2.37 Execute Tanagra 10 37 3.8 Dataset range (including the name of the attributes first row): 12 31 3.95 \$4\$\$1:\$D\$101	1 -
2 38 4.52 0 yes 3 33 2.42 0 yes 4 34 4.5 0 yes 5 31 2.54 0 yes 6 32 9.79 0 yes 7 23 2.5 0 yes 9 29 2.37 Execute Tanagra 10 37 3.8 11 2.8 3.2 11 2.8 3.2 Dataset range (including the name of the attributes first row): 12 13 42 1.98 14 3.0 173	
3 33 2.42 0 yes 4 34 4.5 0 yes 5 31 2.54 0 yes 6 32 9.79 0 yes 7 23 2.5 0 yes 9 29 2.37 Execute Tanagra 10 37 3.8 3.96 11 28 3.2 3.95 13 42 1.98 14 30 173	
4 34 4.5 0 yes 5 31 2.54 0 yes 6 32 9.79 0 yes 7 23 2.5 0 yes 8 28 3.96 0 yes 9 2.9 2.37 Execute Tanagra 10 37 3.8 3.95 11 28 3.2 Dataset range (including the name of the attributes first row): 1 12 31 3.95 1 4.51:\$D\$101 1 13 42 1.98 1 1 1	
5 31 2.54 0 yes 6 32 9.79 0 yes 7 23 2.5 0 yes 8 28 3.96 0 yes 9 2.9 2.37 Execute Tanagra 10 37 3.8 3.95 11 28 3.2 Dataset range (including the name of the attributes first row): 12 31 3.95	
6 32 9.79 0 yes 7 23 2.5 0 yes 8 28 3.96 0 yes 9 29 2.37 Execute Tanagra 10 37 3.8 11 28 3.2 12 31 3.95 13 42 1.98 14 30 173	
7 23 2.5 0 yes 8 28 3.96 0 yes 9 29 2.37 Execute Tanagra 10 37 3.8 11 28 3.2 13 42 1.98 14 30 173	
8 28 3.96 0 vec 9 29 2.37 Execute Tanagra 10 37 3.8 11 28 3.2 12 31 3.95 13 42 1.98 14 30 173	
9 29 2.37 Execute lanagra 10 37 3.8 11 28 3.2 12 31 3.95 13 42 1.98 14 30 173	0.1
10 37 3.8 11 28 3.2 12 31 3.95 13 42 1.98 14 30 173	
11 28 3.2 Dataset range (including the name of the attributes first row): 12 31 3.95 \$A\$1:\$D\$101	
12 31 3.95 13 42 1.96 14 20 1.73	
14 30 173	
OK SM2 Cancel	
16 35 1.91	<u>0</u>
17 41 3.2 0 yes	
18 40 4 0 yes	
19 30 3 1 no	
20 40 10 0 yes	
21 46 3.4 1 no	
22 35 2.35 0 yes	
23 25 1.88 1 no	
24 34 2 0 yes	
25 36 4 1 yes	
20 43 5.14 U yes	
Dråt	

Launching the logistic regression

We add the DEFINE STATUS component into the diagram. We set ACCEPTATION.CREDIT as TARGET, the other attributes as INPUT.

TankaGka 14.33 - [Dataset (tanDr43.txt]) Parameters Analysis Analysis C Age Clear all Clear all Clear all Clear all Clear all Clear all Clear all Clear all Clear all Clear all Clear all Clear all Clear all Clear all Clear all Clear all Clear all Clear all Clear all <
Analysis
Analysis Analysis C Age C Age C Age C Derogatory reports
Analysis Image: Decogatory reports
Dataset (kanSF43:Dxt) Fr43:txt Deceptation Credit Fr43.txt E Clear all Clear selected OK Cancel Help Parameters Acce Acce
Parameters Acce F43.txt F43.txt F43.txt
B Clear all OK OK OK Clear all OK Clear all Clear all Clear all Clear all Clear all OK Clear all OK Clear all Clear all <tr< td=""></tr<>
Image: Clear all Clear selected Image: Clear all Clear selected <t< td=""></t<>
B B Clear all Clear selected OK Cancel Help Parameters Attributes : Target Input
B B Clear all Clear selected OK Cancel Help Parameters Attributes : Target Input
B B Clear all Clear selected OK Cancel Help Parameters Attributes : Target Input
B B Clear all Clear selected OK Cancel Help Parameters Attributes : Target Input
OK Cancel Help Parameters Attributes : Target Input C Age
OK Cancel Help Image: December of the second period Image: December of the second period Parameters Attributes : Target Input Image: December of the second period Adde
Parameters Attributes : Target Input Illustrative C Age
Parameters Attributes : Target Input Illustrative
Attributes : Target Input Illustrative
Target Input Illustrative
G Income per dependent
Data vsualtzation Statist C Derogatory reports Derogatory reports Péature construction
Feature selection Regress Clustering
Spy learning Meta-spy le
Correlation scatterplot
Export dataset Marie data
Lé Scatterplot E View mult

) 📽 🔛 🛛 🗱								
Analysis						Define	status 1	
🖃 🥅 Dataset (tan5F43.tx	t)	Parameters						
Define status 1	Target : 1 Input : 3 Illustrative : 0							
		Results						
		Attri	bute	Target	Input	Illustrat	ive	1.2
		Age		-	yes	-		
		Income.per.depende		t•	yes	-		
		Derogatory	.reports	-	yes	-		
		Acceptatio	n.Credit	yes	•	-		
			Comp	onents				
Data visualization	Statisti	cs	Nonparam	etric st	atistic	s	Instance selection	
Feature construction	Feature sel	ection	tion Regression			Factorial analysis		
PLS	Clusteri	ing	Spv learning		Meta-spv learning			
Spv learning assessment	Scorin	g	Ass	ociatior	ı			
Correlation scatterplot Export dataset Catterplot Scatterplot with label	🔛 View datas 🔁 View multi	set ple scatterpl	ot					

We click on the OK button and we activate the VIEW menu. We obtain the following result.

We add the BINARY LOGISTIC REGRESSION tool (SPV LEARNING tab). We click on the VIEW menu.



The first information displayed is the confusion matrix, computed on the learning sample (CLASSIFIER **PERFORMANCES**). The resubstitution error rate is 0.24. We have the recall and (1.0-precision) for each value of the target attribute.

The MODEL FIT STATISTICS section assesses the global significance of the model. We compare the

Adjustement qu	ality
----------------	-------

Predicted attribute	Acce	ptation.Credit		
Positive value		yes		
Number of examples		100		
Mode	l Fit Statistics			
Criterion	Intercept	Model		
AIC	118.652	108.642		
SC	121.257	119.063		
-2LL	116.652	100.642		
Model Chi ² test (LR)				
Chi-2		16.0094		
d.f.		3		
P(>Chi-2)		0.0011		
	R²-like			
McFadden's R ²		0.1372		
Cox and Snell's R ²		0.1479		
Nagelkerke's R ²		0.2149		

current model to the null model i.e. the model with only the intercept. Roughly speaking, the classifier is relevant if the AIC criterion (or BIC / SC criterion) of the model is lower than the AIC (BIC) of the null model. We observe here that SC(MODEL) = 119.063 < SC(INTERCEPT) = 121.257.

The **MODEL CHI² TEST (LR)** section implements the likelihood ratio test. It allows also to assess the global significance of the model. The CHI-squared statistic CHI-2 = LR = -2LL[INTERCEPT] - (-2LL[MODEL]) = 116.652 - 100.642 = 16.0094. The degree of freedom is equal to the number of explanatory variables (3). Thus, the p-value of the test is 0.0011 according a chi-squared distribution. At the 5% significance level, we conclude that the model is globally significant.

R²-LIKE computes some pseudo r-squared statistics. When the value is close to 0, it means that the model is

not relevant in comparison to the null model.

Attributes in the equation							
Attribute	Coef.	Std-dev	Wald	Signif			
constant	2.745220	1.1430	5.7685	0.0163			
Age	-0.062411	0.0336	3.4541	0.0631			
Income.per.dependent	0.216797	0.1795	1.4587	0.2271			
Derogatory.reports	-1.929304	0.5906	10.6722	0.0011			

Next we obtain the estimated parameters of the model i.e. the coefficients of the logistic regression. Each coefficient is evaluated with the

Wald test. If the p-value is lower than the significance level, the parameter is significant. Here, at 5% significance level, only the parameters for the "Derogatory reports" and the intercept are significant. At the 10% significance level, "Age" becomes significant.

Odds ratios and 95% confidence intervals							
Attribute	Coef.	Low	High				
Age	0.9395	0.8797	1.0034				
Income.per.dependent	1.2421	0.8737	1.7658				
Derogatory.reports	0.1452	0.0456	0.4622				

For people who have the same age and income per dependent, there are 6.89 times more chances to reject the credit when at least one derogatory report is made [1/exp(-1.929304) = 1/0.1452 = 6.89].

These are the **odds-ratios** described into the next table. We obtain also their confidence intervals at 95% confidence level.

Testing the significance of a subset of coefficients

We need to the covariance matrix of the coefficients for the testing the simultaneous significance of a subset of them (Wald test). In Tanagra 1.4.33 version (and later), they are supplied in a second tab (COVARIANCE MATRIX) of the visualization window.

Report		Co	variance matrix		
Cov.Matrix	intercept		Age	Income.per.depende	Derogatory.reports
intercept	1.30645	52	-0.032434781	-0.048171438	-0.077577627
Age	-0.03243	4781	0.0011276876	-0.0015846911	0.0019056894
Income.per.depende	ei-0.04817	1438	-0.0015846911	0.032222176	-0.021162854
Derogatory.reports	-0.07757	7627	0.0019056894	-0.021162854	0.34877701

We can copy these values into the Excel spreadsheet (COMPONENT / COPY RESULTS menu).

📳 CI	lasseur1							×
	А	B	С	D	E	F	G	
1								
2		Cov.Matrix	intercept	Age	Income.per.d	Derogatory.re	ports	
3		intercept	1.31E+00	-3.24E-02	-4.82E-02	-7.76E-02		
4		Age	-3.24E-02	1.13E-03	-1.58E-03	1.91E-03		
5		Income.per.de	-4.82E-02	-1.58E-03	3.22E-02	-2.12E-02		
6		Derogatory.re	-7.76E-02	1.91E-03	-2.12E-02	3.49E-01		
7							rea.	
8								
H 4	► ► Feuil:	L / Feuil2 / Fei	uil3 /)	· //

For instance, if we want to test « H0 : a(AGE) = a(INCOME.PER.DEPENDENT) = 0 », we can compute the test statistic with the following formula

$$W = (-0.062411 \quad 0.216797) \begin{pmatrix} 1.13 \times 10^{-3} & -1.58 \times 10^{-3} \\ -1.58 \times 10^{-3} & 3.22 \times 10^{-2} \end{pmatrix}^{-1} \begin{pmatrix} -0.062411 \\ 0.216797 \end{pmatrix}$$
$$= (-0.062411 \quad 0.216797) \begin{pmatrix} 952.61 & 46.85 \\ 46.85 & 33.34 \end{pmatrix} \begin{pmatrix} -0.062411 \\ 0.216797 \end{pmatrix}$$
$$= 4.0097$$

With the CHI-2 distribution (2 d.f.), the computed p-value is 0.1347. At the 5% significance level, we cannot reject the null hypothesis.

Hosmer and Lemeshow goodness of fit test

The Hosmer and Lemeshow is a test for the overall fit of the model. This test is especially appropriate when the explanatory variables are continuous. It is more robust than the standard chi-square test based on the deviance residuals. The model adequately fits the data if the test highlights non-significance. We add the HOSMER LEMESHOW TEST component (SPV LEARNING ASSESSMENT tab) behind the regression. This is the only location where we can insert it into the diagram anyway.

We click on the VIEW menu. We obtain the table for the calculations. The value of the test statistic is CHI-2 = 4.4530 with the p-value = 0.8141. Because the p-value is higher than the significance level (5%), we conclude than the model fits adequately the observed dataset.



Reliability diagram

The reliability diagram compares the observed probability and the estimated posterior probability of the model. The instances are subdivided into groups. The model fits the dataset if the points are aligned on a straight line.



First, we must compute the posterior probability of "yes" supplied by the model. We use the SCORE (SCORING tab). We set "yes" as the positive class value.

We click on the VIEW menu to launch the calculations. A new column is added to the current dataset. It corresponds to the computed posterior probability to be "yes" for each instance.

Then, we add the DEFINE STATUS tool. We set ACCEPTATION.CREDIT as TARGET, and SCORE_1 (the SCORE computed before) as INPUT. We note that we can set many scores as input. It is useful when we want to compare the performance of classifiers. We note also that the score column is not necessarily a probability. It must only make possible to rank the examples according to their tendency to be positive.



Then we add the RELIABILITY DIAGRAM tool (SCORING tab). We click on the PARAMETRES menu. We specify the positive class value (ACCEPTATION.CREDIT = "yes").

We observe that we can build the reliability diagram on the selected examples (learning sample) or on the unselected examples (test sample). If the dataset was subdivided previously using the SAMPLING component for instance, we can compute the reliability diagram on the test sample. The result is more robust.

	onent window Heip					
Ar	nalysis		Scoring 1			
∃ 🧰 Dataset (tan5F43.b	(t)		Parameters			
Define status 1	earning 1 (Binary logistic reg	Positive class value : yes	Positive class value : yes			
⊢? Hosmer ⊢î↓ Scoring ⊢î. Defin	Lemeshow Fest 1 1 e status 2 elfability Diagram 1	Compu Compu Compu	1 inplicately by			
	Parameters Execute View	Positive cla Used ex © Unsel	ass value : yes ted lected OK	Cancel Help		
D		Components	1	-		
Data visualization	Statistics	Nonparametric statistics	Instance selection	Feature construction		
Feature selection	Regression	Factorial analysis	PLS	Clustering		
	Moto cov loorning	Sov learning assessment	Scoring	Association		

We click on the VIEW menu.



The model is not very good. We observe that the scores are overestimated into the second group; they are underestimated into the third one.

ROC curve

The ROC curve is a very useful tool for assessing the performance of a classifier. Among their multiple interpretations, we will say that it allows to evaluate the ability of the model to assign a higher score to the positive instances (than the negative instances). We add the ROC CURVE tool (SCORING). We set the following parameters.

A Dataset (tan5F43.t Define status 1 Define st	nalysis xt) Learning 1 (Binary logistic rej Lemeshow Test 1 1	Report Diag Scoring curve Parameters	gram		1
Dataset (tan5F43.t	xt) Learning 1 (Binary logistic rej Lemeshow Test 1 1	Scoring curve	1.		
	ne status 2 Reliability Diagram 1 Roc curve 1 Parameters Execute View	Positive class va Used examp Selected Unselected	lue: yes		•
			ОК	Cancel	Help
Data visualization Feature selection Spv learning	Statistics Regression Meta-spv learning	Nonparametric statistics Pactorial analysis Spy learning ascessment	Instance selection PLS Scoring	Featu	re construction Clustering Association
Lift curve Posterior Prob	Precision-Recall curve	Roc curve			

We click on the VIEW menu. We obtain the ROC curve.



The comparison of the models is also possible with this tool. We obtain the area under curve AUC = 0.7575. It seems that our model is "fair" (http://gim.unmc.edu/dxtests/ROC3.htm).

By comparing the results provided by several indicators, we can get a more reliable opinion about the quality of the model. This aspect is very important. We must not focus on a single indicator.

Residual analysis

The residual analysis allows to check the validity of the model. It allows also to detect the outliers and influential points. We add the LOGISTIC REGRESSION RESIDUALS tool (SPV LEARNING ASSESSMENT tool) behind the logistic regression. We click on the VIEW menu.

TANAGRA 1.4.33 - [Logist	tic Regression Residuals 1]	- Constraint and the second		lgonigi			
🍸 File Diagram Component Window Help 📃 🗗 🗙							
An	alysis	Report Re	esiduals				
Dataset (tan5F43.txt)		ha	at_logReg_1	pearson_logReg_1	std_pearson_logReg_	1 difcł 🔺	
🗐 🚼 Define status 1		1	0.024189	0.508257	0.5145	18	
Supervised Learning 1 (Binary Logistic regu		2	0.017029	0.545989	0.55069	98	
6 Hermor Lemorbow Test 1		3	0.019377	0.449589	0.4540	10	
		4	0.015076	0.506325	0.51018	36	
□ ↓ Scoring 1		5	0.078039	0.238053	0.24792	23	
🖻 🚰 Define status 2		6	0.022197	0.396182	0.40065	53	
🖳 🖉 Reliability Diagram 1		7	0.017131	0.395297	0.39872	27	
Roc curve 1		8	0.016699	0.484538	0.48863	36	
Logistic Regression Residuals 1		9	0.01/845	0.532631	0.53/44	18	
		10	0.014700	0.429242	0.43243	32	
		11	0.015140	0.434561	0.43/88	39	
		12	0.056369	-1.318656	-1.35/4	/1	
		13	0.024359	-1.866338	-1.88949	33	
		14	0.016119	0.480355	0.4842	/3	
		15	0.02/664	0.614183	0.6228		
	4					Þ	
		Components					
Data visualization	Statistics	Nonparametric statistics	Instance sele	ection Feature	Feature construction		
Feature selection	Regression	Factorial analysis	PLS	Clustering			
Spv learning	Meta-spv learning	Spv learning assessment	Scoring	g Association			
18 Bias-variance decomposition _7 Hosmer Lemeshow Test							
Bootstrap ?Leave-One-Out ?Train-test							
Cross-validation 🗮 Logistic Regression Residuals							
						Lat all asit	

Some indicators are available. They are computed for each individual.

- HAT: leverage;
- PEARSON: Pearson residual;
- STD_PEARSON: standardized Pearson residual;
- DIFCHISQ: contribution to the Pearson statistic;
- DEVIANCE: deviance residual;
- STD_DEVIANCE: standardized deviance residual;
- DIFDEV: contribution to the deviance;
- COOK: Cook's distance ;
- DEFBETA: DFBETA for each explanatory variable, including the intercept;
- DFBETAS: standardized DFBETA.

The values which are higher than (and/or lower than) the cut values are in bold. But it is often more interesting to copy the values into a spreadsheet and sorting the table according some indicators. We can better detect the suspicious examples.

Into the REPORT tab, we have a summary of the number of outliers or influential points detected for each indicator.

An	alusis	Report	Re	siduals		
Dataset (tan5F43.txt) Define status 1 Supervised Learning 1 (Binary logistic regiled) Josephine Status 2 Josephine Status 2 Z Reliability Diagram 1 Z Roc curve 1 Logistic Regression Residuals 1			1007		Bosultr	
		Residuals Regressic Criterion Hat (Leverage)	Residuals and Inlfuential Points - Logistic Regression Cut-off values Criterion Lower Upper Count Hat (Leverage) - 0.0800 12			
		DIFCHISQ DIFDEV Cook's D DFBETAS	- - -0.2000	3.8400 3.8400 0.0417 0.2000	7 4 5 35	
Data visualization eature construction PLS ov learning assessment	Statistics Feature selection Clustering Scoring	Components Nonparametric statistics Regression Spv learning Association			nstance Factorial Meta-spv	selection analysis learning
Bias-variance decompos Bootstrap Cross-validation	tion _? Hosmer Lemeshov ? Leave-One-Out Logistic Regressio	w Test ? Test ? Trai on Residuals	: n-test			

4 - Analysis in R

We do not detail all operations with R. We give only the commands that allow to achieve the requested results. The source code "LOGISTIC_REGRESSION_DIAGNOSTICS.R" is included into the archive which is distributed with this tutorial.

Importing the data file and performing the logistic regression

We set the following commands to import the data file and to perform the logistic regression. We load the dataset in the XLS (Excel) file format using the "xlsReadWrite" package.

🥜 D:\DataMining\Databases_for_mining\dataset_for_soft_dev_and_comparison\logistic regression\residuals\logistic_regress	sio 🗖 🗖 💌
#clear the internal memory	^
<pre>rm(list=ls())</pre>	E
#in order to handle a XLS file format	
library(xlsReadWrite)	
#loading the dataset	
setwd("D:/DataMining/Databases_for_mining/dataset_for_soft_dev_and_comparison/ld	ogistic regre
<pre>donnees <- read.xls(file = "logistic_regression_diagnostics.xls",rowNames = FALS</pre>	SE, sheet=1)
summary (donnees)	
#performing the logistic regression	
<pre>modele <- glm(Acceptation.Credit ~ ., data = donnees, family = "binomial")</pre>	
resume <- summary (modele)	
print (resume)	-
<	∎ 1

We obtain the same results as Tanagra.

```
R Console
                                                           - • •
> print(resume)
Call:
glm(formula = Acceptation.Credit ~ ., family = "binomial", data = do$
Deviance Residuals:
            1Q Median 3Q
7108 0.5680 0.7034
   Min
                                      Max
-2.2621 -0.7108
                                  2.0814
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
                     2.74522 1.14306 2.402 0.01632 *
-0.06241 0.03358 -1.858 0.06311 .
(Intercept)
Age
                    -0.06241
Income.per.dependent 0.21680 0.17951 1.208 0.22714
Derogatory.reports -1.92930 0.59058 -3.267 0.00109 **
____
Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
(Dispersion parameter for binomial family taken to be 1)
   Null deviance: 116.65 on 99 degrees of freedom
Residual deviance: 100.64 on 96 degrees of freedom
AIC: 108.64
Number of Fisher Scoring iterations: 4
•
```

To obtain the prediction of the model and the confusion matrix, we set.



We obtain.

```
R Console
> #computing the confusion matrix
> prediction <- ifelse (predict (modele, type="response") > 0.5, "yes", "no")
> print (table (donnees$Acceptation.Credit, prediction))
    prediction
        no yes
        no 8 19
        yes 5 68
>
```

Covariance matrix of coefficients

We need the covariance matrix in order to test the significance of a subset of coefficients. It is associated to the "summary object".

R Console				
<pre>> #obtaining the cova > print(attributes(re \$names</pre>	riance matrix sume))			*
[1] "call"	"terms"	"family"	"deviance"	"aic"
[6] "contrasts"	"df.residual"	"null.deviance"	"df.null"	"iter"
[11] "deviance.resid"	"coefficients"	"aliased"	"dispersion"	"df"
<pre>[16] "cov.unscaled"</pre>	"cov.scaled"			
<pre>\$class [1] "summary.glm" > print(resume\$cov.un</pre>	scaled)			
	(Intercept)	Age Income.pe	r.dependent Derog	gatory.reports
(Intercept)	1.30658434 -0.03	2438283 -	0.048175586	-0.077548615
Age	-0.03243828 0.00	1127781 -	0.001584562	0.001904885
Income.per.dependent	-0.04817559 -0.00	1584562	0.032222092	-0.021163524
Derogatory.reports	-0.07754862 0.00	1904885 -	0.021163524	0.348782597
•		m		

R has powerful tools for the matrix operations. We can use them directly.

Hosmer - Lemeshow test

For the Hosmer-Lemeshow test, the ROC curve and the reliability diagram, we need the "score" column. We use the **predict(.)** command.



Then, we write a function to compute the Hosmer and Lemeshow statistic. I think this source code is very simplistic, but it allows to the reader to understand the details of the treatments³.

```
D:\DataMining\Databases_for_mining\dataset_for_soft_dev_and_comparison\logistic regression\residuals\logistic_regression_...
fanother Hosmer-Lemeshov procedure
ffrom https://stat.ethz.ch/pipermail/r-help/2008-September/174044.html
cut.score <- cut(score, breaks = quantile(score, probs = seq(0, 1, 0.1)), include.lowest = T)
obs <- xtabs(cbind(1 - y, y) ~ cut.score)
expect <- xtabs(cbind(1 - score, score) ~ cut.score)
chisq <- sum((obs - expect)^2/expect)
P <- 1 - pchisq(chisq, 10 - 2)
print(c("X^2" = chisq, Df = 10 - 2, "P(>Chi)" = P))
```

³ For instance, here is a sorter source code loaded from the web. Of course, we obtain the same results.

```
🥜 D:\DataMining\Databases_for_mining\dataset_for_soft_dev_and_comparison\logistic regression\residuals\logistic_reg... 👝 💷 📧
  #Hosmer - Lemeshow test
  y <- ifelse(unclass(donnees$Acceptation.Credit)==2, 1, 0)
  total <- 0.0
  previous <- 0.0
  for (p in seq(0.1, 1, 0.1)) {
    #covered examples into the quantile
    seuil <- quantile(score,p)</pre>
    examples <- (score > previous & score <= seuil)
     #number of covered examples
    m.obs <- length(which(examples))</pre>
     #positive
    m.pos.obs <- sum(y[examples])</pre>
    m.pos.expected <- sum(score[examples])</pre>
    #negative
    m.neg.obs <- m.obs - m.pos.obs
    m.neg.expected <- m.obs - m.pos.expected
     #statitic
    total <- total + (m.pos.obs - m.pos.expected) ^2/m.pos.expected
    total <- total + (m.neg.obs - m.neg.expected) ^2/m.neg.expected
     #next
    previous <- seuil
  3
  print(c("Hosmer Lemeshow Statistic" = total, "p-value" = pchisg(total,8,lower.tail=F)))
< III
```

The result is consistent with those of Tanagra.

Residual analysis and influential points

Some commands enable to compute the indicators for the residual analysis.



influence() provides the leverage (hat values), DFBETA, deviance and Pearson residuals.

Influence.measures() provides DFBETAS, DFFIT, COVRATIO, Cook's distance, leverage.

Other commands are available. We observe that all the results are consistent with those of Tanagra, excepting the DFBETA(s). I have not found the origin of the differences. I note only that the DFBETA(s) provided by Tanagra, SAS and SPSS (state-of-the-art commercial tools) are identical.

5 - Conclusion

In this tutorial, we tried to give an overview of the tools used to assess and diagnose the logistic regression. Some are specific to the regression (Hosmer-Lemeshow test, analysis of residuals), while others are more generic, they can be used for any classifier which can provide a prediction or a "score" (confusion matrix, ROC curve).