Chapter 19
Linear, Logistic, and Cox Regression for Outcome Prediction with Unpaired Data (20, 55, and 60 Patients)

General Purpose

To assess whether linear, logistic and Cox modeling can be used to train clinical data samples to make predictions about groups and individual patients.

Specific Scientific Question

How many hours will patients sleep, how large is the risk for patients to fall out of bed, how large is the hazard for patients to die.

This chapter was previously published in "Machine learning in medicine-cookbook 1" as Chap. 4, 2013.

T.J. Cleophas, A.H. Zwinderman, *Machine Learning in Medicine - a Complete Overview*, DOI 10.1007/978-3-319-15195-3_19

Linear Regression, the Computer Teaches Itself to Make Predictions

Var 1	Var 2	Var 3	Var 4	Var 5
0,00	6,00	65,00	0,00	1,00
0,00	7,10	75,00	0,00	1,00
0,00	8,10	86,00	0,00	0,00
0,00	7,50	74,00	0,00	0,00
0,00	6,40	64,00	0,00	1,00
0,00	7,90	75,00	1,00	1,00
0,00	6,80	65,00	1,00	1,00
0,00	6,60	64,00	1,00	0,00
0,00	7,30	75,00	1,00	0,00
0,00	5,60	56,00	0,00	0,00
1,00	5,10	55,00	1,00	0,00
1,00	8,00	85,00	0,00	1,00
1,00	3,80	36,00	1,00	0,00
1,00	4,40	47,00	0,00	1,00
1,00	5,20	58,00	1,00	0,00
1,00	5,40	56,00	0,00	1,00
1,00	4,30	46,00	1,00	1,00
1,00	6,00	64,00	1,00	0,00
1,00	3,70	33,00	1,00	0,00
1,00	6,20	65,00	0,00	1,00

Var 1 = treatment 0 is placebo, treatment 1 is sleeping pill
Var 2 = hours of sleep
Var 3 = age
Var 4 = gender
Var 5 = comorbidity

SPSS 19.0 is used for analysis, with the help of an XML (eXtended Markup Language) file. The data file is entitled “linoutcomeprediction” and is in extras.springer.com. Start by opening the data file.

Command:

Click Transform….click Random Number Generators….click Set Starting Point ….click Fixed Value (2000000)….click OK….click Analyze….Regression…. Linear….Dependent: enter hoursofsleep….Independent: enter treatment and age…. click Save….Predicted Values: click Unstandardized….in XML Files click Export final model….click Browse….File name: enter "exportlin"….click Save….click Continue….click OK.

Coefficients[a]						
Model		Unstandardized coefficients		Standardized coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	,989	,366		2,702	,015
	Treatment	−,411	,143	−,154	−2,878	,010
	Age	,085	,005	,890	16,684	,000

[a]Dependent variable: hoursofsleep

The output sheets show in the coefficients table that both treatment and age are significant predictors at $p<0.10$. Returning to the data file we will observe that SPSS has computed predicted values and gives them in a novel variable entitled PRE_1. The saved XML file will now be used to compute the predicted hours of sleep in 4 novel patients with the following characteristics. For convenience the XML file is given in extras.springer.com.

Var 1	Var 2	Var 3	Var 4	Var 5
,00	6,00	66,00	,00	1,00
,00	7,10	74,00	,00	1,00
,00	8,10	86,00	,00	,00
,00	7,50	74,00	,00	,00

Var 1 = treatment 0 is placebo, treatment 1 is sleeping pill
Var 2 = hours of sleep
Var 3 = age
Var 4 = gender
Var 5 = comorbidity

Enter the above data in a new SPSS data file.

Command:

Utilities….click Scoring Wizard….click Browse….click Select….Folder: enter the exportlin.xml file….click Select….in Scoring Wizard click Next….click Use value substitution….click Next….click Finish.

The above data file now gives individually predicted hours of sleep as computed by the linear model with the help of the XML file.

Var 1	Var 2	Var 3	Var 4	Var 5	Var 6
,00	6,00	66,00	,00	1,00	6,51
,00	7,10	74,00	,00	1,00	7,28
,00	8,10	86,00	,00	,00	8,30
,00	7,50	74,00	,00	,00	7,28

Var 1 = treatment 0 is placebo, treatment 1 is sleeping pill
Var 2 = hours of sleep
Var 3 = age
Var 4 = gender
Var 5 = comorbidity
Var 6 = predicted hours of sleep

Conclusion

The module linear regression can be readily trained to predict hours of sleep both in groups and, with the help of an XML file, in individual future patients.

Note

More background, theoretical and mathematical information of linear regression is available in Statistics applied to clinical studies, 5th edition, Chaps. 14 and 15, entitled "Linear regression basic approach" and "Linear regression for assessing precision, confounding, interaction", pp 161–176 and 177–185, Springer Heidelberg Germany 2012, from the same authors.

Logistic Regression, the Computer Teaches Itself to Make Predictions

Var 1	Var 2	Var 3	Var 4	Var 5
,00	1,00	50,00	,00	1,00
,00	1,00	76,00	,00	1,00
,00	1,00	57,00	1,00	1,00
,00	1,00	65,00	,00	1,00
,00	1,00	46,00	1,00	1,00
,00	1,00	36,00	1,00	1,00
,00	1,00	98,00	,00	,00
,00	1,00	56,00	1,00	,00
,00	1,00	44,00	,00	,00
,00	1,00	76,00	1,00	1,00
,00	1,00	75,00	1,00	1,00
,00	1,00	74,00	1,00	1,00
,00	1,00	87,00	,00	,00

Var 1 department type
Var 2 falling out of bed (1 = yes)
Var 3 age
Var 4 gender
Var 5 letter of complaint (1 = yes)

Only the first 13 patients are given, the entire data file is entitled "logoutcomeprediction" and is in extras.springer.com.

SPSS 19.0 is used for analysis, with the help of an XML (eXtended Markup Language) file. Start by opening the data file.

Command:

Click Transform….click Random Number Generators….click Set Starting Point ….click Fixed Value (2000000)….click OK….click Analyze….Regression …. Binary Logistic….Dependent: enter fallingoutofbed ….Covariates: enter departmenttype and letterofcomplaint….click Save….in Predicted Values click Probabilities….in Export model information to XML file click Browse…. File name: enter "exportlog"….click Save….click Continue….click OK.

Variables in the equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Departmenttype	1,349	,681	3,930	1	,047	3,854
	Letterofcomplaint	2,039	,687	8,816	1	,003	7,681
	Constant	−1,007	,448	5,047	1	,025	,365

[a]Variable(s) entered on step 1: departmenttype, letterofcomplaint

In the above output table it is shown that both department type and letter of complaint are significant predictors of the risk of falling out of bed. Returning to the data file we will observe that SPSS has computed predicted values and gives them in a novel variable entitled PRE_1. The saved XML file will now be used to compute the predicted hours of sleep in 5 novel patients with the following characteristics. For convenience the XML file is given in extras.springer.com.

Var 1	Var 2	Var 3	Var 4	Var 5
,00	,00	67,00	,00	,00
1,00	1,00	54,00	1,00	,00
1,00	1,00	65,00	1,00	,00
1,00	1,00	74,00	1,00	1,00
1,00	1,00	73,00	,00	1,00

Var 1 department type
Var 2 falling out of bed (1 = yes)
Var 3 age
Var 4 gender
Var 5 letter of complaint (1 = yes)

Enter the above data in a new SPSS data file.

Command:

Utilities….click Scoring Wizard….click Browse….click Select….Folder: enter the exportlog.xml file….click Select….in Scoring Wizard click Next….mark Probability of Predicted Category….click Next….click Finish.

The above data file now gives individually predicted probabilities of falling out of bed as computed by the logistic model with the help of the XML file.

Var 1	Var 2	Var 3	Var 4	Var 5	Var 6
,00	,00	67,00	,00	,00	,73
1,00	1,00	54,00	1,00	,00	,58
1,00	1,00	65,00	1,00	,00	,58
1,00	1,00	74,00	1,00	1,00	,92
1,00	1,00	73,00	,00	1,00	,92

Var 1 department type
Var 2 falling out of bed (1 = yes)
Var 3 age
Var 4 gender
Var 5 letter of complaint (1 = yes)
Var 6 Predicted Probability

Conclusion

The module binary logistic regression can be readily trained to predict probability of falling out of bed both in groups and, with the help of an XML file, in individual future patients.

Note

More background, theoretical and mathematical information of binary logistic regression is available in Statistics applied to clinical studies 5th edition, Chaps. 17, 19, and 65, entitled "Logistic and Cox regression, Markov models, Laplace transformations", "Post-hoc analyses in clinical trials", and "Odds ratios and multiple regression", pp 199–218, 227–231, and 695–711, Springer Heidelberg Germany 2012, from the same authors.

Cox Regression, the Computer Teaches Itself to Make Predictions

Var 1	Var 2	Var 3	Var 4
1,00	1,00	,00	65,00
1,00	1,00	,00	66,00
2,00	1,00	,00	73,00
2,00	1,00	,00	91,00

Var 1	Var 2	Var 3	Var 4
2,00	1,00	,00	86,00
2,00	1,00	,00	87,00
2,00	1,00	,00	54,00
2,00	1,00	,00	66,00
2,00	1,00	,00	64,00
3,00	,00	,00	62,00
4,00	1,00	,00	57,00
5,00	1,00	,00	85,00
6,00	1,00	,00	85,00

Var 1 follow up in months
Var 2 event (1 = yes)
Var 3 treatment modality
Var 4 age

Only the first 13 patients are given, the entire data file is entitled "Cox outcome prediction" and is in extras.springer.com.

SPSS 19.0 is used for analysis, with the help of an XML (eXtended Markup Language) file. Start by opening the data file.

Command:

Click Transform….click Random Number Generators….click Set Starting Point ….click Fixed Value (2000000)….click OK….click Analyze….Survival….Cox Regression….Time: followupmonth….Status: event….Define event: enter 1…. Covariates: enter treatment and age….click Save….mark: Survival function…. In Export Model information to XML file click Browse…. File name: enter "export-Cox"….click Save….click Continue….click OK.

Variables in the equation

	B	SE	Wald	df	Sig.	Exp(B)
Treatment	–,791	,332	5,686	1	,017	,454
Age	,028	,012	5,449	1	,020	1,028

In the above output table it is shown that both treatment modality and age are significant predictors of survival. Returning to the data file we will now observe that SPSS has computed individual probabilities of survival and gave them in a novel variable entitled SUR_1. The probabilities vary from 0.00 to 1.00. E.g., for the first patient, based on follow up of 1 month, treatment modality 0, and age 65, the computer has computed a mean survival chance at the time of observation of 0.95741 (= over 95 %). Other patients had much less probability of survival. If you would have limited sources for further treatment in this population, it would make sense not to burden with continued treatment those with, e.g., less than 20 % survival

probability. We should emphasize that the probability is based on the information of the variables 1, 3, 4, and is assumed to be measured just prior to the event, and the event is not taken into account here.

Var 1	Var 2	Var 3	Var 4	SUR_1
1,00	1,00	,00	65,00	,95741

The saved XML file will now be used to compute the predicted probabilities of survival in 5 novel patients with the following characteristics. For convenience the XML file is given in extras.springer.com. We will skip the variable 2 for the above reason.

Var 1	Var 2	Var 3	Var 4
30,00		1,00	88,00
29,00		1,00	67,00
29,00		1,00	56,00
29,00		1,00	54,00
28,00		1,00	57,00

Var 1 follow up in months
Var 2 event (1 = yes)
Var 3 treatment modality
Var 4 age

Enter the above data in a new SPSS data file.

Command:

Utilities….click Scoring Wizard….click Browse….click Select….Folder: enter the exportCox.xml file….click Select….in Scoring Wizard click Next….mark Predicted Value….click Next….click Finish.

The above data file now gives individually predicted probabilities of survival as computed by the Cox regression model with the help of the XML file.

Var 1	Var 2	Var 3	Var 4	Var 5 PredictedValue
30,00		1,00	88,00	,18
29,00		1,00	67,00	,39
29,00		1,00	56,00	,50
29,00		1,00	54,00	,51
28,00		1,00	57,00	,54

Var 1 follow up in months
Var 2 event (1 = yes)
Var 3 treatment modality
Var 4 age
Var 5 predicted probability of survival (0.0–1.0)

Conclusion

The module Cox regression can be readily trained to predict probability of survival both in groups and, with the help of an XML file, in individual future patients. Like outcome prediction with linear and logistic regression models, Cox regression is an important method to determine with limited health care sources, who of the patients will be recommended expensive medications and other treatments.

Note

More background, theoretical and mathematical information of binary logistic regression is available in Statistics applied to clinical studies 5th edition, Chaps. 17 and 31, entitled “Logistic and Cox regression, Markov models, Laplace transformations”, and “Time-dependent factor analysis”, pp 199–218, and pp 353–364, Springer Heidelberg Germany 2012, from the same authors.