GENERALISED LINEAR MODELS. AN INTRODUCTORY APPLICATION TO LOGISTIC REGRESSION WITH DICHOTOMOUS RESPONSE

Preface



Dear Ph.D. students, here you can find the first basic exercise for exploring generalized linear models.

Some basic questions will focus on the model structure, so as to choose the appropriate model within the GLIMs family and to read and diagnose the model fit.

Let us consider file 1 (<u>http://www.csun.edu/~ata20315/psy524/main.htm</u>).

I have converted files from the format .sav into .csv since we won't use Spss much, at least in the next future. In reality it was not strictly necessary, for all software considered can read .sav and several other formats easily. However, spreadsheet are very versatile and easily recognised.

I will use Jamovi first, since it is may be used by other colleagues. Then I will sketch the same problem in Jasp, so as to allow you to have an idea of other user friendly and open source software. I will also provide you the basic instruction for R (I will skip the screenshots in Rstudio).

Here are the variables in file1.csv., with labels, categories included, when needed. Score means that the variable is a score of a psychological test.

gender o= male, 1=female; ethn: Ethnicity - o=minority; 1=majority sos sexual opinion survey - score ego egotism score n neuroticism score e extraversion score o openness score a agreeableness score c conscientiousness score disoi dichotomized sexual openness and intimacy o=closed, 1=open

<u>Aim</u>: We want to model sexual openness and intimacy. Which GLIM do we use? We need to focus on the response type, disoi, dichotomous. Therefore, we use logistic regression.

Before estimating the model

We will act as we had already checked for data quality, for outliers etc.

What assumptions do we need to check? Absence of multicollinearity (no predictor is a linear combination – i.e. a linear transform- of one or more other predictors).

Another important assumption is independence errors, i.e. a between subject design. When facing mixed models (multilevel models, hierarchical models etc.), you will see the violation of this assumption and its consequences.

Remember to check that the software has set the correct variables scales. In our case, gender, ethnicity and sexual openness must be declared nominal. All the others are quantitative as they are scores. <u>Modelling</u>

Predict disoi by everything else. Check significance levels. Ask odds ratios and measure of model fit. Interpret. What do the 'reference levels' mean? By default, they are all set to o.

<u>Comment</u>

Comment on the results and note doubts.

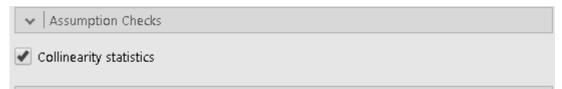
Jamovi

Nomir	nal '	•						
	ominal rdinal				Sps	s.	is the equivalent of s	scale in
🧼 Ca	ontinuo	us			In ti	ruth is c	juantitative	
🥒 ID								
= (Data Ana	alyses	Edit			jamovi	- file1	
Exploration	₹ T-Tests	₽ АNOVA		quencies Factor				
				DATA	VARIABLE			
				gen				
					ription ure type 🔒 Nom	ninal 🔻	Levels	
					ype Integer	Tillar •	0	- L
					ig values		1	
							Retain unused levels	5
🔒 gend			< sos	🔶 ego	🤣 n	< e	Description	
1 2	1	0	89 52	162.5 151.0	26	^	Results	
3	0	0	52	185.5	38			
4	1	0	57	179.5	28			

Specification of our model

	Edit			
	Con		Dependent Variable	
		\rightarrow	🐣 disoi	<u>_</u>
	Regression Frequencies		Covariates	
	Correlation Matrix	\rightarrow	< sos	A
Э	Partial Correlation		🔶 ego	
1	Linear Regression		🧼 n	
С	-		🔶 e	
С	Logistic Regression		A	 T
7			Factors	
2	2 Outcomes Binomial	\rightarrow	🐣 ethnic	
1	N Outcomes		🐣 gender	
C	Multinomial			
2	Ordinal Outcomes			
÷				

We can check for multicollinearity only after the model specification, otherwise the package does not know the predictors we are interested in.



Model fit measures. At lesson we mentioned some pseudo R measures, AIC, here we require also BIC (Bayesian Information Criterion) and overall model fit.

✓ Model Fit	
Fit Measures	Pseudo R ²
🕑 Deviance	McFadden's R ^z
💽 AIC	Cox & Snell's R ^z
🖌 BIC	✓ Nagelkerke's R ²
🕑 Overall model test	

Then we need estimated coefficient. Give your own selection or try the different ones.

✓ Model Coefficients	
Omnibus Tests	Odds Ratio
Likelihood ratio tests	Odds ratio
Estimate (Log Odds Ratio) Confidence interval Interval 95 %	Confidence interval

Jasp

Jasp reads directly many formats, but not xlxs, then I converted into .cs.

Fixing scales is quick.



We choose logistic regression from the regression menu.

Here we have chosen the model.

Logistic Regression	0	000
	Dependent Variable 🦺 disoi	
	Method	h 🔥
	Enter	V
	Covariates	
	SOS	
	📏 ego	
	N n	
	N e	
	> 0	
	Factors	
	👶 gender	
	👶 ethnic	
		🥌 📶 📗

Here we choose measure of model fit and also diagnostics. Please note that some measures included here are in a specific (prediction) section in Jamovi.

 Statistics 	
Descriptives	Performance Diagnostics
Factor descriptives	Confusion matrix
	Proportions
Regression Coefficients	Performance Metrics
🕜 Estimates	AUC
From 5000 bootstraps	Sensitivity / Recall
Standardized coefficients	Specificity
Odds ratios	Precision
Confidence intervals	F-measure
Interval 95.0 %	Brier score
Odds ratio scale	H-measure
Robust standard errors	
📃 Vovk-Sellke maximum p-ratio	
Residuals	
Casewise diagnostics	
Standard residual > 3	
Cook's distance > 1	
) All	

In the same menu, you can choose some plots. In Jamovi you can find some in the Prediction submenu.

	🗐 file1*	(G:\II mio Drive\dida	ttica\dottorato\s	cuoladottorato,	_202021)										-		×			
	Ξ	Descriptives	T-Tests		Mixed Models		Frequencies										+	۶ ۱	<mark>Զ</mark> Condi	idi
ľ				•	Covariates			Π	T	Results								*		^
					 ego n e o Factors 		\			Logistic I		sion								
				•	🥚 🙏 gen					Model	Devianc		BIC	df	X²	р	Mo			
					🦔 eth	TC .				H _o H ₁	390.135 314.055				76.079	< .001	1			
							h, &			Coefficients										
		Model															Walc			
		Statistics											Estimate	Standard Error	z	Wald S	tatistic			
1	1	Diots							•	(In (tJaspC	olumn5.	Encoded)	-0.924	2.492 0.009	-0.371 4.572		0.137 0.906			
										sos ego			0.041	0.009	4.572		1.678			
		antial Plots			Residual Plots					n			-0.015	0.022	-0.707		0.500			
		Display conditional est	timates plots		Predicted	 residual plot 				e			0.045	0.028	1.637		2.680			
		Confidence interval 9	5.0 %		Predictor	- residual plots				0			-0.017	0.026	-0.663		0.439			
		Show data points			Coursed f	Pearson residuals	olot			a			-0.082	0.025	-3.209		0.296 4.582			
		onow outo points					plot			gender (1)			0.450	0.318	1,414		1.999			
	Resid	ual Type			Performance Plo	ots				ethnic (1)			-0.438	0.308	-1.423		2.024			
	0	Deviance			ROC plot					Note. disoi le	evel "l' code	d as class 1.								
		Pearson			Cutoff ste	p 0.2			:											
					Add c	utoff labels			0											
					PR plot															
4					_															120%
					Cutoff ste													10:5	54	
ł	۶ 🗄	O Scrivi qui per es	eguire la rice	rca	0	Hi 🐠	单 🔕	22			U		🖻 🙆 🕼	0 🌡 🏭 🛥	😵 d») 💗	8	J. 0		/2020 C	$\overline{}$

R Studio

For GLIM in R I will refer to:

- Faraway, J. J. (2016). Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models. CRC press.
- <u>https://stats.idre.ucla.edu/other/dae/</u>

Personally, I use R studio, I am told the same for our PhD school.

gender (double) 👻	ethnic (double) –	sos (numeric) 👻	ego (numeric) 🔻	n (numeric) *	e (numeric) 👻	o (numeric) 🔻	a (numeric) 🔻	€ (numeric) ▼	disoi (double) 👻
1	0	89	162.5	26	35	36	36	51	0
0	1	52	151.0	38	36	34	40	33	0
0	0	57	185.5	32	43	43	52	48	0
1	0	57	179.5	28	46	31	35	48	1
1	0	90	167.0	42	53	47	37	31	0
1	0	87	157.0	28	37	35	51	57	0
0	1	59	176.5	47	41	40	44	36	0
0	0	57	152.5	32	38	39	46	45	0
0	0	72	215.0	36	3/	50	22	/0	0

In importing the data file, I checked the measurement scale.

Here I paste the syntax, please remember that in Rstudio some commands in our case ¾ of what Ishow you, use dialog boxes. No need to endless writing.

To simplify: commands in the red box use dialog boxes; The hash mark tells us that the line is a comment commands in the green box need writing, in our case the logistic model; (minimal) output in the light blue box.

- + "numeric", "numeric", "numeric", "numeric",
- + "numeric", "numeric", "numeric", "numeric",
- + "numeric"))

#I save the imported file in R data format: file1.Rdata save.image("G:/II mio Drive/didattica/dottorato/scuoladottorato_202021/PhD2020_exercises/file1.RData")

#modello logistico. Installo glm2, da R studio si usano finestre di dialogo
library(glm2)
#chiamo I modello logistico mylogit
mylogit <- glm(disoi ~ gender + ethnic + sos + ego + n + e + o + a + c, data = file1, family = "binomial")
chiedo le stime del modello
summary(mylogit)</pre>

Call:

 $glm(formula = disoi \sim gender + ethnic + sos + ego + n + e + o + a + c, family = "binomial", data = file1)$

Deviance Residuals:

Min 1Q Median 3Q Max

 $-2.2380 \ -0.8648 \ -0.4838 \ \ 0.9502 \ \ 2.9000$

Coefficients:

Estimate Std. Error z value Pr(>|z|) (Intercept) -0.923674 2.492353 -0.371 0.71093 0.449905 0.318249 1.414 0.15745 gender ethnic -0.438290 0.308061 -1.423 0.15481 sos 0.041233 0.009018 4.572 4.82e-06 *** 0.013945 0.010764 1.296 0.19514 ego -0.015343 0.021695 -0.707 0.47943 n 0.045171 0.027591 1.637 0.10160 е -0.017373 0.026221 -0.663 0.50760 0 -0.081744 0.025475 -3.209 0.00133 ** а -0.046491 0.021718 -2.141 0.03231 * С

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 (Dispersion parameter for binomial family taken to be 1) Null deviance: 390.13 on 286 degrees of freedom Residual deviance: 314.06 on 277 degrees of freedom

AIC: 334.06

Number of Fisher Scoring iterations: 4



Please feel free to discuss this assignment together and then contact me.