7.2 Logistic Regression - Working with data

Tuesday, 08 November 2022 12:56

Summary	• Logistic Regression
	X_1 = Academic performance during the undergraduate degree. X_2 = Academic performance during the MBA. X_3 = Industry experience prior to joining the MBA program. X_4 = Participation in the co-curricular and extra-curricular activities. $Y = 1$ if the student gets placed, and zero otherwise.
	Predicting the placements • Since the problem has been reduced to predicting the value of Y using X_1, X_2, X_3 and X_4, is this regression? • Can these attributes be used to predict whether a student will pick up a job during the placement process? • Answer is yes! Through "Logistic regression". • However, we need to pay attention to our response variable. • Since the response variable is binary (or generically speaking, categorical), we can't use the regular regression method and expression. • Logistic regression is used to predict the dependent categorical variable.
• How do we solve this problem? • $Odds$ (of success) = ?	**Solution method: Regression** • If this was modeled as a multiple linear regression, we would have $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$ • Since, our Y is binary, assumptions of the regression model won't hold and we won't get good predictions. • Can we try using probabilities? • That is: $Pr\{Y=1\}$ as a predictor. Then, our response variable has values between 0 and 1. • However, if we calculate ODDs, then we can get out of these limits. $$Odds(success\ in\ placements) = \frac{\Pr(Y=1)}{\Pr(Y=0)}$$ • If $P(Y=1) = 0.9$ and $P(Y=0) = 0.1$, then we say the odds of success is $9:1$.
• How do we use $Odds$ in regression equation? • From there, how do you calculate $P(Y=1)$?	**Solution method: Regression** • More commonly, Log values are used. That is, Log of the odds. • As a result, we have: (dropping the error term) $$Log(Odds) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$ $$Odds = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4}$$ $$\Pr(Y=1) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4}}$$ • Here, we're assuming that the log has a base of e. • Let $Odds = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2 + \beta_4 X_4} = A$ $$Odds = \frac{P(Y=1)}{P(Y=0)} = \frac{P(Y=1)}{1 - P(Y=1)} = A$$ $$\Rightarrow A - A\,P(Y=1) = P(Y=1)$$ $$\Rightarrow A = P(Y=1)(1+A)$$ $$\Rightarrow P(Y=1) = \frac{A}{1+A} = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2 + \beta_4 X_4}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2 + \beta_4 X_4}}$$
	Solution method: Regression

- Now we can run the regression model and estimate the regression coefficients (the β's).
- The objective function used for this estimation: maximization of the log-likelihood. That is the log of probability of the correct prediction.

- See the Excel sheet.

This is the correlation matrix:

Correlation Matrix

	MBA CGPA	*Experience*	*UG CGPA*	*Extra-curricular*	*Day-0 placed*
MBA CGPA	1				
Experience	-0.038867107	1			
UG CGPA	0.348301526	0.170294352	1		
Extra-curricular	0.16846311	-0.070032269	0.176627455	1	
Day-0 placed	0.599259002	0.166540606	0.424267443	0.354241475	1

- We're interested in Correlation of the response variable (Day-0 placed) with the other explanatory variables.
- Also, if you notice the correlation coefficients among the explanatory variables, the correlations don't seem to be strong except between MBA CGPA and UG CGPA.

This is the working of Logistic Regression

b0	b1	b2	b3	b4
-41.7512	3.2741492	0.5915958	0.83093	0.87624

SUM of Log-Likelihood	-7.875726483
Cutoff	0.5

					Observed Y				Likelihood		Predicted Y
Student	MBA CGPA	Experience	UG CGPA	Extra-curricular	Day-0 placed	Logit - Log(odds)	Odds	Prob of Day-0 job	Prob of correct estimate	Log-Likelihood	Classification
1	9.1	2.3	8.1	8.6	1	3.67042506	39.26859	0.975166751	0.975166751	-0.025146796	1
2	8.9	0	8.7	8.9	1	2.41635488	11.20494	0.918065973	0.918065973	-0.085486025	1
3	7	3.9	8	5.13	0	-5.38238078	0.004597	0.00457583	0.99542417	-0.004586331	0
4	9.1	1.1	7.8	4.9	0	-0.5308569	0.588101	0.370317052	0.629682948	-0.462538843	0
5	8.2	0.7	9.3	9.13	1	1.2386607	3.450988	0.775330804	0.775330804	-0.254465497	1
6	6.5	1.5	7.9	4.2	0	-9.3372815	8.81E-05	8.80708E-05	0.999911929	-8.80747E-05	0

? No idea how we got the b values.
I ran MLR on the data and the values I got are not matching with given values.

	Coefficients
Intercept	-3.058911577
MBA CGPA	0.241765018
Experience	0.06389908
UG CGPA	0.098114914
Extra-curricular	0.08658737

b0	b1	b2	b3	b4		
-41.7512	3.2741492	0.5915958	0.83093	0.87624		
Student	MBA CGPA	Experience	UG CGPA	Extra-curricular	Day-0 placed	Logit - Log(odds)
1	9.1	2.3	8.1	8.6	1	3.67042506
2	8.9	0	8.7	8.9	1	2.41635488
3	7	3.9	8	5.13	0	-5.38238078
4	9.1	1.1	7.8	4.9	0	-0.5308569
5	8.2	0.7	9.3	9.13	1	1.2386607

Logit-Log (odds)
$= b0 + (b1 \times MBA\ CGPA) + (b2 \times Experience) + (b3 \times UG\ CGPA) + (b4 \times Extracurricular)$

=$A$2+SUMPRODUCT($B$2:$E$2, B6:E6)

Basically this is what we're computing here:
$$Log(Odds) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

- Understand the working

Logit - Log(odds)	Odds
3.67042506	39.26859
2.41635488	11.20494
-5.38238078	0.004597
-0.5308569	0.588101

$$Odds = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4}$$

- Odds = $e^{Logit-Log(odds)}$

Odds	Prob of Day-0 job
39.26859	0.975166751
11.20494	0.918065973
0.004597	0.00457583
0.588101	0.370317052

$$\Pr(Y=1) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4}}$$

$$Pr(Y=1) = \frac{Odds}{1 + Odds}$$

Day-0 placed	Prob of Day-0 job	Prob of correct estimate
1	0.975166751	0.975166751
1	0.918065973	0.918065973
0	0.00457583	0.99542417
0	0.370317052	0.629682948

Prob of correct estimate =
If Day-0 placed = 1
Prob of Day−0 job
Else
1 − Prob of Day−0 job

Likelihood	
Prob of correct estimate	Log-Likelihood
0.975166751	-0.025146796
0.918065973	-0.085486025
0.99542417	-0.004586331
0.629682948	-0.462538843

Log-likelihood = ln(Prob of correct estimate)

SUM of Log-Likelihood	-7.875726483
Cutoff	0.5

Then we sum up all the values in Log-Likelihood column – and that becomes our objective function.

The objective function used for this estimation: maximization of the log-likelihood. That is the log of probability of the correct prediction.