Logistic Regression Exercises

Ethan Marcano

2/25/2022

Exercise 1: Predicting the Baseball World Series Champion

A) Exploring the dataset

Table 1: A brief look at the data

Team	League	Year	RS	RA	W	OBP	SLG	BA	RankSeason	RankPlayoffs	NumCompetitors	WonWorldSeries
NYY	AL	1962	817	680	96	0.337	0.426	0.267	2	1	2	1
SFG	NL	1962	878	690	103	0.341	0.441	0.278	1	2	2	0
LAD	NL	1963	640	550	99	0.309	0.357	0.251	2	1	2	1
NYY	AL	1963	714	547	104	0.309	0.403	0.252	1	2	2	0
NYY	AL	1964	730	577	99	0.317	0.387	0.253	1	2	2	0
STL	NL	1964	715	652	93	0.324	0.392	0.272	2	1	2	1

i) Each row of the dataset represent's a playoff's team's performance in a particular year. Through the years, different numbers of teams have been invited to the playoffs. How has the number of teams making it to the playoffs changed, according to this dataset?

- The number of competitors has increased from two teams at the beginning to 10 teams in the present day.
- Teams invited to the playoffs are somewhat lower in the regular season because of the the number of competitors invited.

ii) Given that a team has made it to the playoffs, it is much harder to win the World Series if there are 10 teams competing for the championship versus just two. Therefore we have the variable NumCompetitors in our dataset. NumCompetitors contains the number of total teams making the playoffs in the year of the observation. For instance, NumCompetitors is 2 for the 1962 New York Yankees, but it is 8 for the 1998 Boston Red Sox. Without knowing anything else about the teams in the playoffs, can you think of a simple model that uses NumCompetitors to predict the probability of a team winning?

- A linear regression model using NumCompetitors shows that it is statistically significant in determining the probability of a team winning.

Table 2: Summary of Linear Regression

term	estimate	std.error	statistic	p.value
(Intercept)	0.4355346	0.0723201	6.022317	0.0000000
NumCompetitors	-0.0389937	0.0109108	-3.573864	0.0004245

B) Building a logistic regression model to predict the winner

i) When we are not sure which of our variables are useful in predicting a particular outcome, it is often helpful to build bivariate models, which are models that predict the outcome using a single independent variable. Build a bivariate logistic regression model using each of the following variables as the independent variable to predict WonWorldSeries and the entire dataset as the training set each time: Year, RS, RA, W, OBP, SLG, BA, RankSeason, NumCompetitors, and League. You should have created 10 logistic regression models. Describe each of the models by giving the regression equation and the accuracy of the model. For which models is the independent variable significant? In your opinion, which are the best models and why?

- The year, number of competitors, RankSeason, and RA are all significant in their individual models.

ii) Now, build a logistic model using all of the variables that you found to be significant in the bivariate models as the independent variables, and the entire dataset to train the model. Are all of the independent variables significant in this model? Why would some independent variables be significant in the bivariate model using that variable, but then not significant in a model that uses more than one independent variables? Be sure to provide numerical evidence for your claim.

- As none of the p-values are significant, none of the independent variables that were originally found to be significant are significant in the multivariate model. This may be due to colinearity.

	Dependent variable:
	WonWorldSeries
NumCompetitors	−0.025
	(0.028)
Year	−0.001
	(0.004)
RA	−0.0002
	(0.0004)
RankSeason	−0.008
	(0.016)
Constant	3.436
	(8.478)
Observations	244
Log Likelihood	-113.568
Akaike Inf. Crit.	237.136

Note: *p<0.1; **p<0.05; ***p<0.01

iii) Using any number of the independent variables that you found to be significant in the bivariate models, find what you think is the best model, and justify why you think it is the best. How many independent variables are used in your final model?

iv) Do your findings in this problem confirm or reject the claim that the playoffs is more about luck than skill? Why?

- Yes, it confirms that the claim. The most significant variable, NumCompetitors, has no bearing on a team's skill.