Logistic Regression

Cody Frisby

10/28/2017

Training the logistic model in R with a 75% random sample of the original data frame, we get the model represented in the table below.

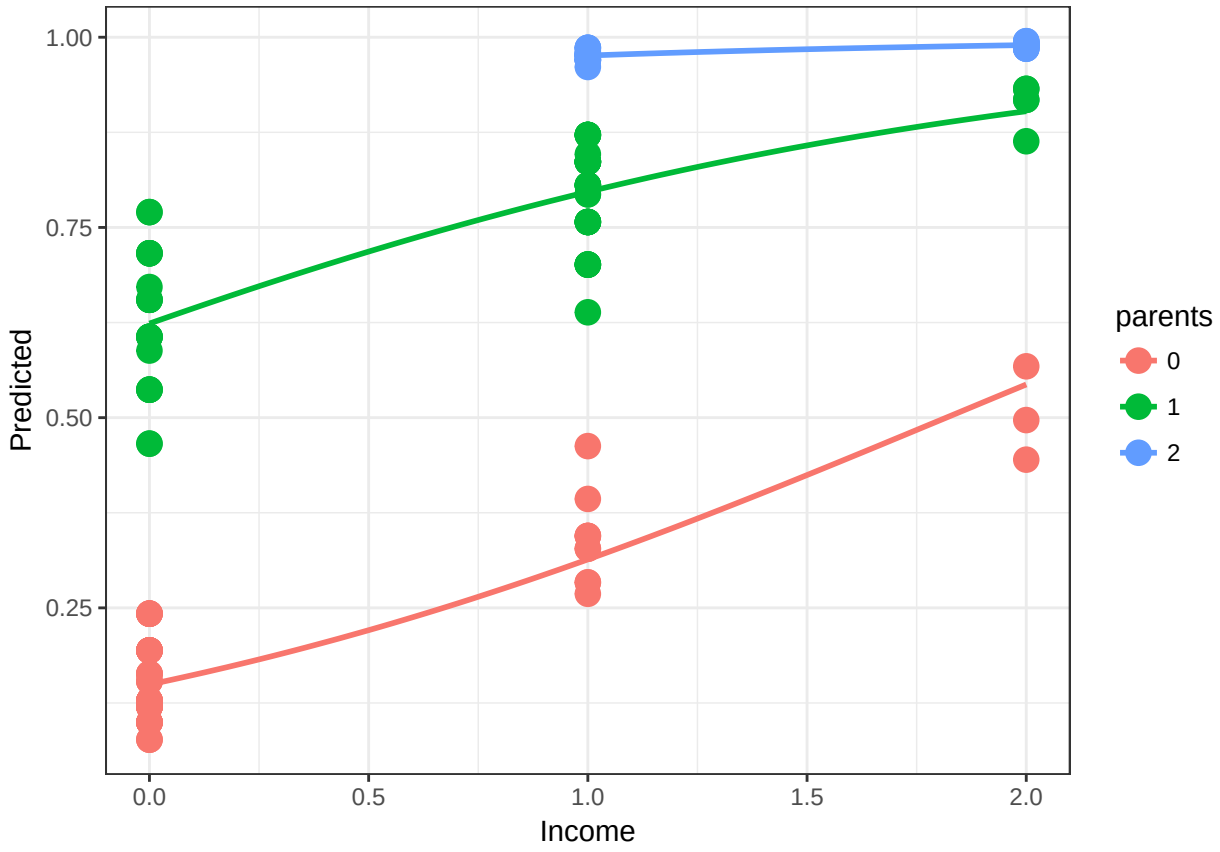
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.0543199	0.9625063	-3.173299	0.0015072
Parent_Grad	2.6332170	0.5176833	5.086540	0.0000004
Gender	0.4932951	0.4393730	1.122725	0.2615543
Income_Level	0.9898878	0.3705572	2.671349	0.0075547
Num_Siblings	0.2842631	0.2251915	1.262317	0.2068347

Whether or not one or both parents graduated from college or not appears to be the strongest predictor of graduation (**Parent_Grad**). **Gender** and **Num_Siblings** do not appear to have much influence on graduation. **Income_Level** appears to have some influence on graduation with a *p-value* of 0.0075547.

The table below is what's often called a *confusion matrix*. Essentially, it's a comparison between the observed response variable and the model's prediction of that variable. Here I compare a simple random sample of 25% of the original data with the models predictions. The diagonal of the matrix shows strong agreement between observed **0** and predicted **0** as well as for **1**. It would appear that we have a miss-classification error of 0.0545455.

	0	1
0	26	0
1	3	26

When we have few enough variables in our model, I think it's interesting to visualize it.



As we can see (and this would be much better if income was continuous) as income increases the prediction of graduation also increases. We also have three different distinct lines where neither parent graduated (0), one graduated (1), or both graduated (2). Interestingly, where both parents graduated (the blue line) it appears that the probabilities do not stray far from 1.