**Wesbrook dataset. Logistic regression.**

First check variables summary:

A close up of a sign

Description automatically generated

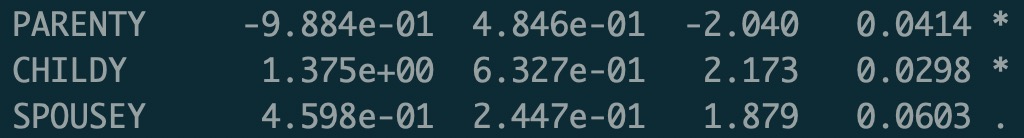
Delete those variables which have more than 50% missing values

Nullifying these variables, “ID”, “INDUPDT” (Date of last update of personal data) and “BIGBLOCK” (one level factor variable)

Building model, getting summary table and leaving only statistically significant variables:

A picture containing object

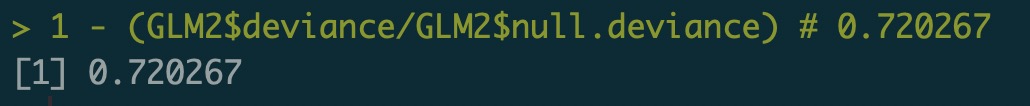
Description automatically generated





Even considering that SPOUSEY (Spouse of UBC student or alumnus code) and OTHERACTY (Participation in other activities) have very low significance I still prefer to leave them because I think that they may be useful.

Now, after excluding lot of variables constructing new model and getting McFadden value:



Next step - transform numerical variables into Log to create non-linear model, creating model gives such McFadden value:

A close up of a logo

Description automatically generated

Since there is only 1 numerical variable left in the model – there is no possibility in building mixed model, so as a conclusion I can say that comparing 2 models – non-linear performs a little bit better than linear according to McFadden value. It’s not surprise because that variable (“TOTLGIVE”) which was transformed into Log form is the most statistically significant.